Nonparametric estimation of composite functions

Anatoli B. Juditsky, Oleg Lepski, Alexandre B. Tsybakov

Nonparametric estimation of composite functions

Anatoli B. Juditsky[1], Oleg V. Lepski[2] and Alexandre B. Tsybakov[3]

[1)] Laboratoire Jean Kuntzmann
Université Grenoble 1, B.P. 53, 38041 Grenoble, France
e-mail: anatoli.juditsky@imag.fr
[2)] Laboratoire d'Analyse, Topologie et Probabilités
Université de Provence, 39, rue F.Joliot Curie, 13453 Marseille, France
e-mail: lepski@cmi.univ-mrs.fr
[3)] Laboratoire de Probabilités et Modèles Aléatoires
Université Paris 6, 4, pl.Jussieu, Case 188, 75252 Paris, France
e-mail: tsybakov@ccr.jussieu.fr

May 21, 2007

Abstract

We study the problem of nonparametric estimation of a multivariate function $g : \mathbb{R}^d \to \mathbb{R}$ that can be represented as a composition of two unknown smooth functions $f : \mathbb{R} \to \mathbb{R}$ and $G : \mathbb{R}^d \to \mathbb{R}$. We suppose that f and G belong to some known smoothness classes of functions and we construct an estimator of g which is optimal in a minimax sense for the sup-norm loss. The proposed methods are based on aggregation of linear estimators associated to appropriate local structures, and the resulting procedures are nonlinear with respect to observations.

Keywords *multidimensional nonparametric estimation, minimax estimation, adaptive estimation, composite functions, single index model*

1 Introduction

In this paper we study the problem of nonparametric estimation of an unknown function $g : \mathbb{R}^d \to \mathbb{R}$ in the multidimensional gaussian white noise model described by the stochastic differential equation

$$X_\varepsilon(dt) = g(t)dt + \varepsilon W(dt), \quad t = (t_1, \dots, t_d) \in \mathcal{D} \tag{1}$$

where $\mathcal{D}$ is an open interval in $\mathbb{R}^d$ containing $[-1, 1]^d$, W is the standard Brownian sheet in $\mathbb{R}^d$ and $0 < \varepsilon < 1$ is a known noise level. Our goal is to estimate the function g on the set $[-1, 1]^d$ from the observation $\{X_\varepsilon(t),\ t \in \mathcal{D}\}$. For $d = 2$ this corresponds to the problem of image reconstruction from observations corrupted by additive noise. We consider observation set $\mathcal{D}$ which is larger than $[-1, 1]^d$ in order to avoid the discussion of boundary effects.

To measure the performance of estimators, we will use the risk function determined by the sup-norm $\|\cdot\|_\infty$ on $[-1, 1]^d$: for $g : \mathbb{R}^d \to \mathbb{R}$, $0 < \varepsilon < 1$, $p > 0$, and for an arbitrary

estimator $\tilde{g}_\varepsilon$ based on the observation $\{X_\varepsilon(t),\ t \in \mathcal{D}\}$ we consider the risk

$$R_\varepsilon\big(\tilde{g}_\varepsilon, g\big) = \mathbb{E}_g\Big(\|\tilde{g}_\varepsilon - g\|_\infty^p\Big).$$

Here and in what follows $\mathbb{E}_g$ denotes the expectation w.r.t. the distribution $\mathbb{P}_g$ of the observation $\{X_\varepsilon(t),\ t \in \mathcal{D}\}$ satisfying (1).

We will suppose the $g \in \mathcal{G}_\mathsf{s}$, where $\{\mathcal{G}_\mathsf{s},\ \mathsf{s} \in \mathsf{S}\}$ is a collection of functional classes indexed by $\mathsf{s} \in \mathsf{S}$. The functional classes $\mathcal{G}_\mathsf{s}$ that we will consider consist of *smooth composite functions* and below we discuss in detail this choice.

For a given class $\mathcal{G}_\mathsf{s}$ we define the maximal risk

$$R_\varepsilon\big(\tilde{g}_\varepsilon, \mathcal{G}_\mathsf{s}\big) = \sup_{g \in \mathcal{G}_\mathsf{s}} R_\varepsilon\big(\tilde{g}_\varepsilon, g\big). \tag{2}$$

Our first aim is to study the asymptotics, as the noise level ε tends to 0, of the minimax risk

$$\inf_{\tilde{g}_\varepsilon} R_\varepsilon\big(\tilde{g}_\varepsilon, \mathcal{G}_\mathsf{s}\big)$$

where $\inf_{\tilde{g}_\varepsilon}$ denotes the infimum over all estimators of g. We suppose that parameter s is known, and therefore the functional class $\mathcal{G}_\mathsf{s}$ is fixed. In other words, we are interested in minimax estimation of g. We find the minimax rate of convergence $\phi_\varepsilon(\mathsf{s})$ on $\mathcal{G}_\mathsf{s}$, i.e., the rate which satisfies $\phi_\varepsilon^p(\mathsf{s}) \asymp \inf_{\tilde{g}_\varepsilon} R_\varepsilon\big(\tilde{g}_\varepsilon, \mathcal{G}_\mathsf{s}\big)$ and we construct an estimator attaining this rate, called rate optimal estimator in asymptotic minimax sense. The estimator depends on parameter s which restricts its application in practice. We discuss approaches to treat adaptation to s and state some conjectures on this issue. We suggest a possible construction of such an adaptive procedure. Further details will be given in a forthcoming paper.

2 Motivation

It is well known that the main difficulty in estimation of multivariate functions is the curse of dimensionality: the best attainable rate of convergence of the estimators decreases very fast as the dimension grows. To illustrate this effect, suppose, for example, that the underlying function g belongs to $\mathcal{G}_\mathsf{s} = \mathbb{H}_d(\alpha, L)$, $\mathsf{s} = (\alpha, L)$, $\alpha > 0, L > 0$, where $\mathbb{H}_d(\alpha, L)$ is an isotropic Hölder class of functions. We give the exact definition of this functional class later. Here we only mention that $\mathbb{H}_d(\alpha, L)$ consists of functions g with bounded partial derivatives of order $\leq \lfloor \alpha \rfloor$ and such that, for all $x, y \in \mathcal{D}$,

$$|g(y) - P_g(x, y - x)| \leq L\|x - y\|^\alpha,$$

where $P_g(x, y - x)$ is the Taylor polynomial of order $\leq \lfloor \alpha \rfloor$ obtained by expansion of g around the point x, and $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$. Parameter α characterizes the isotropic (i.e., the same in each direction) smoothness of function g.

If we use the risk (2), uniformly on $\mathbb{H}_d(\alpha, L)$ the rate of convergence of estimators cannot be asymptotically better than

$$\psi_{\varepsilon,d}(\alpha) = \big(\varepsilon\sqrt{\ln(1/\varepsilon)}\big)^{2\alpha/(2\alpha+d)}$$

(cf. (9; 10; 23; 20; 5)). This is also the minimax rate on $\mathbb{H}_d(\alpha, L)$: it is achieved, for example, by a kernel estimator with properly chosen bandwidth and kernel. More results on asymptotics of the minimax risks in estimation of multivariate functions can be found

in (11; 12; 2; 3). It is clear that if α is fixed and d is large enough this asymptotics is too pessimistic to be used for real data: the value $\psi_{\varepsilon,d}(\alpha)$ is small only if the noise level ε is unreasonably small. On the other hand, if the noise level ε is realistically small and α is small the above asymptotics can be of no use already in dimensions 2 or 3.

At the origin of this phenomenon is the fact that the d-dimensional isotropic Hölder class $\mathbb{H}_d(\alpha, L)$ is too massive in terms of its metric entropy. To "overcome" the curse of dimensionality one usually considers models with slimmer functional classes (i.e., classes with smaller metric entropy). There are several ways to do it.

- A first way is to impose a restriction on the smoothness parameter of the functional class. For the class $\mathbb{H}_d(\alpha, L)$, a convenient restriction is to assume that the smoothness α increases with the dimension, and thus the class becomes smaller (its metric entropy decreases). For instance, we can suppose that $\alpha = \kappa d$ with some fixed $\kappa > 0$. Then the dimension disappears from the expression for $\psi_{\varepsilon,d}(\alpha)$, which means that we escape from the curse of dimensionality. However, the condition $\alpha = \kappa d$ or other similar restrictions linking smoothness and dimension are usually difficult to motivate. One interesting related example seems to be the class of functions with absolutely integrable multivariate Fourier transform (1).

- Another way is to impose a *structural assumption* on the function g to be estimated. Two classical examples are provided by the single index and additive structures (cf., e.g., (24; 6; 8)).

  The *single index structure* is defined by the following assumption on g: there exist a function $F_0 : \mathbb{R} \to \mathbb{R}$ and a vector $\vartheta \in \mathbb{R}^d$ with $\|\vartheta\| = 1$ such that $g(x) = F_0(\vartheta^T x)$.

  The *additive structure* is defined by the following assumption: there exist functions $F_i : \mathbb{R} \to \mathbb{R},\ i = 1, \dots, d$, such that $g(x) = F_1(x_1) + \cdots + F_d(x_d)$, where x_j is the j-th component of $x \in \mathbb{R}^d$.

  If we suppose that $F_i \in \mathbb{H}_1(\alpha, L),\ i = 0, \dots, d$, then in both cases function g can be estimated with the rate $\left(\varepsilon\sqrt{\ln(1/\varepsilon)}\right)^{2\alpha/(2\alpha+1)}$ which does not depend on the dimension and coincides with the minimax rate $\psi_{\varepsilon,1}(\alpha)$ of estimation of functions on $\mathbb{R}$.

In general, under structural assumptions the rate of convergence of estimators improves, as compared to the slow d-dimensional rate $\psi_{\varepsilon,d}(\alpha)$. For the above examples the rate does not depend on the dimension.

However, it is often quite restrictive to assume that g has some simple structure, such as the single index or additive one, *on the whole domain of its definition*. In what follows we refer to this assumption as *global structure*.

A more flexible way of modeling is to suppose that g has a *local structure*. For instance, we can assume that g is well approximated by some single index or additive structure (or by a combination both) in a small neighborhood of a given point x. Local structure depends on x and remains unchangeable within the neighborhood. Such an approach can be used to model much more complex objects than the global one. However, the form of the d-dimensional neighborhood and the local structure should be chosen by the statistician in advance, which makes the local approach rather subjective.

In the present paper we try to find a compromise between the global and local modeling. Our idea is to consider a sufficiently general global model which would generate suitable local structures, and thus would allow us to construct estimators with nice statistical properties. Such a global model should satisfy the following requirements.

(i) The optimal performance of estimators under this model is better than the performance of methods based only on the smoothness properties of the underlying function.

(ii) The model contains "global" parameters such that their choice automatically generates interpretable local structures and associated domains of localization (neighborhoods). Adaptation to such global parameters, if it is feasible, means adaptation to different local models simultaneously.

(iii) The model contains a tuning parameter such that its choice allows one to reduce the influence of the dimension.

We argue that this program can be realized for global models where the underlying function g is a composition of two smooth functions.

3 Smooth composite functions

We now define our global structural model. We will assume that g is a *composite function*, i.e., that $g(t) = f(G(t))$ for all $t \in \mathbb{R}^d$ where $f : \mathbb{R} \to \mathbb{R}$ and $G : \mathbb{R}^d \to \mathbb{R}$ with $d \geq 2$.

We will further suppose that f and G are smooth functions such that $f \in \mathbb{H}_1(\gamma, L_1)$ and $G \in \mathbb{H}_d(\beta, L_2)$ where γ, L_1, β, L_2 are positive constants. Here and in what follows $\mathbb{H}_1(\gamma, L_1)$ and $\mathbb{H}_d(\beta, L_2)$ are the Hölder class on $\mathbb{R}$ and the isotropic Hölder class on $\mathbb{R}^d$ respectively (see Definition 1 below). The class of composite functions g with such f and G will be denoted by $\mathbb{H}(\mathcal{A}, \mathcal{L})$, where $\mathcal{A} = (\gamma, \beta) \in \mathbb{R}_+^2$ and $\mathcal{L} = (L_1, L_2) \in \mathbb{R}_+^2$.

This model is a generalization of the single index model: instead of the linear function we have here a general $G(\cdot)$.

The performance of an estimation procedure will be measured by the sup-norm risk (2) where we set $\mathsf{s} = (\mathcal{A}, \mathcal{L})$ and $\mathcal{G}_\mathsf{s} = \mathbb{H}(\mathcal{A}, \mathcal{L})$. The global parameter of the model is $\mathsf{s} = (\mathcal{A}, \mathcal{L})$, and we will show that the choice of $\mathcal{A}$ leads to different local structures. Note also that the value of s determines the quality of estimation associated to our model, i.e., the rate of convergence of the minimax risk.

We start with the following definitions:

Definition 1. *Fix $\alpha > 0$ and $L > 0$. Let $\lfloor \alpha \rfloor$ be the largest integer which is strictly less than α, and for $\vec{k} = (k_1, \ldots, k_d) \in \mathbb{N}^d$ set $|\vec{k}| = k_1 + \cdots + k_d$. The isotropic Hölder class $\mathbb{H}_d(\alpha, L)$ is the set of all functions $G : \mathbb{R}^d \to \mathbb{R}$ having on $\mathbb{R}^d$ all partial derivatives of order $\lfloor \alpha \rfloor$ and such that*

$$\sum_{0 \leq |\vec{k}| \leq \lfloor \alpha \rfloor} \sup_{x \in \mathbb{R}^d} \left| \frac{\partial^{|\vec{k}|} G(x)}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}} \right| \leq L,$$

$$\left| G(y) - \sum_{0 \leq |\vec{k}| \leq \lfloor \alpha \rfloor} \frac{\partial^{|\vec{k}|} G(x)}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}} \prod_{j=1}^{d} \frac{(y_j - x_j)^{k_j}}{k_j!} \right| \leq L \|y - x\|^\alpha, \quad \forall\, x, y \in \mathbb{R}^d. \tag{3}$$

where x_j and y_j are the jth components of x and y.

Definition 2. *Let $\mathcal{G}$ be some class of functions on $\mathbb{R}^d$. We say that $\mathcal{G}$ is a class of effective smoothness $\alpha > 0$ if $\mathcal{G} \subseteq \mathbb{H}_d(\alpha, L)$ for some $L > 0$, and $\mathcal{G} \not\subseteq \mathbb{H}_d(\alpha', L')$ for all $\alpha' > \alpha$, $L' > 0$.*

Observe that for any $\alpha > 0$, $L > 0$, the effective smoothness of the single index and additive classes is equal to α. Next, it is easy to see that for any $\mathcal{A} = (\gamma, \beta) \in \mathbb{R}^2_+$ and $\mathcal{L} = (L_1, L_2) \in \mathbb{R}^2_+$ the effective smoothness of $\mathbb{H}(\mathcal{A}, \mathcal{L})$ is

$$\alpha_{\gamma,\beta} \triangleq \begin{cases} \gamma\beta & \text{if } 0 < \gamma, \beta \le 1, \\ \min(\gamma, \beta) & \text{otherwise.} \end{cases} \tag{4}$$

We say that there is an improvement in statistical performance of estimators due to the structure if the minimax rate of convergence associated to the class $\mathcal{G}$ is $o(\psi_{\varepsilon,d}(\alpha))$, as $\varepsilon \to 0$, where α is the effective smoothness of $\mathcal{G}$. In other words, the knowledge of a structure allows certain improvement if the best estimator, based only on the smoothness properties, converges slower than the best estimator which takes into account the whole structure. For example, for the classes of functions with single-index or additive structure there is always an improvement due to the structure, because the corresponding minimax rate is $\left(\varepsilon\sqrt{\ln(1/\varepsilon)}\right)^{2\alpha/(2\alpha+1)} = o(\psi_{\varepsilon,d}(\alpha))$.

For the class $\mathbb{H}(\mathcal{A}, \mathcal{L})$ of composite functions this property is not always true. For certain values of $\mathcal{A} = (\gamma, \beta)$ no improvement due to the structure can be expected. This happens if our structural assumption is essentially equivalent to the fact that g belongs to some isotropic Hölder class of functions of full dimension d, and the knowledge of the composition structure does not help to improve the statistical analysis. Such an effect appears in the following two zones of (γ, β).

1°. *Zone of slow rate*: $0 < \gamma, \beta \le 1$.

Clearly, in this zone $\mathbb{H}(\mathcal{A}, \mathcal{L}) \subset \mathbb{H}_d(\gamma\beta, L_3)$, where L_3 is a positive constant depending only on γ, β and $\mathcal{L}$. Due to this inclusion a standard kernel estimator with properly chosen bandwidth and the boxcar kernel converges with the rate $\psi_{\varepsilon,d}(\gamma\beta) = \left(\varepsilon\sqrt{\ln(1/\varepsilon)}\right)^{2\gamma\beta/(2\gamma\beta+d)}$. It is not hard to see (cf. Section 8) that this rate is optimal, i.e., that a lower bound on the minimax risk holds with the same "slow" rate $\psi_{\varepsilon,d}(\gamma\beta)$ (note that $\gamma\beta \le 1$). As the effective smoothness of $\mathbb{H}(\mathcal{A}, \mathcal{L})$ for $0 < \gamma, \beta \le 1$ equals to $\gamma\beta$, there is no improvement due to the structure.

2°. *Zone of inactive structure*: $\gamma \ge \beta,\ \gamma \ge 1$.

In this zone we easily get the inclusions $\mathbb{H}_d(\beta, L_4) \subset \mathbb{H}(\mathcal{A}, \mathcal{L}) \subset \mathbb{H}_d(\beta, L_5)$, where L_4 and L_5 are positive constants depending only on β and $\mathcal{L}$. To show the left inclusion it suffices to fix a linear function f. Therefore, the asymptotics of the minimax risk on $\mathbb{H}(\mathcal{A}, \mathcal{L})$ is the same as for any isotropic Hölder class $\mathbb{H}_d(\beta, \cdot)$. In particular, a standard kernel estimator converges with the rate $\psi_{\varepsilon,d}(\beta)$. Note that here we estimate as if there were no structure, and the asymptotics of the minimax risk does not depend on γ. This explains why we refer to this zone as that of *inactive structure*.

We finally remark that if $\beta \le 1$ the composite function g is rather nonsmooth. The effective smoothness equals to $(1 \wedge \gamma)\beta$, and in view of the above discussion, the minimax rate of convergence of estimators on $\mathbb{H}(\mathcal{A}, \mathcal{L})$ is the same as on the Hölder class $\mathbb{H}_d((1 \wedge \gamma)\beta, \cdot)$. This is a very slow rate $\psi_{\varepsilon,d}((1 \wedge \gamma)\beta)$. Therefore, only for $\beta > 1$ one can expect to find estimators with interesting statistical properties.

4 Main results

In this section we state the main results and outline the estimation method. The formal description of the estimation procedure and the proofs are deferred to Sections 6 and 8 – 9 respectively.

4.1 Lower bound for the risks of arbitrary estimators

For any $\mathcal{A} = (\gamma, \beta) \in \mathbb{R}_+^2$ define

$$\pi(\mathcal{A}) = \frac{2\gamma}{2\gamma + 1 + (d-1)/\beta} \wedge \frac{2}{2 + d/\beta} \wedge \frac{2}{2 + d/(\gamma\beta)}, \tag{5}$$

and

$$\phi_\varepsilon(\gamma, \beta) = \left(\varepsilon\sqrt{\ln(1/\varepsilon)}\right)^{\pi(\mathcal{A})}. \tag{6}$$

In an expanded form, we may write

$$\phi_\varepsilon(\gamma, \beta) = \begin{cases} \left(\varepsilon\sqrt{\ln(1/\varepsilon)}\right)^{\frac{2\gamma}{2\gamma+1+(d-1)/\beta}} & \text{if} \quad \beta > 1, \beta \geq d(\gamma-1)+1, \\ \left(\varepsilon\sqrt{(\ln 1/\varepsilon)}\right)^{\frac{2}{2+d/\beta}} & \text{if} \quad \gamma > 1, \beta < d(\gamma-1)+1, \\ \left(\varepsilon\sqrt{\ln(1/\varepsilon)}\right)^{\frac{2}{2+d/(\gamma\beta)}} & \text{if} \quad (\gamma, \beta) \in (0,1]^2. \end{cases} \tag{7}$$

The boundaries between the zones of these three different rates in $\mathbb{R}_+^2$ are presented by the dashed lines in Figure 1.

An asymptotic lower bound for the minimax risk on $\mathbb{H}(\mathcal{A}, \mathcal{L})$ is given by the following theorem.

Theorem 1. *For any $\mathcal{A} = (\gamma, \beta) \in \mathbb{R}_+^2$ and any $p > 0$ we have*

$$\liminf_{\varepsilon\to 0} \inf_{\tilde{g}_\varepsilon} \sup_{g \in \mathbb{H}(\mathcal{A},\mathcal{L})} \mathbb{E}_g \left[\left(\phi_\varepsilon^{-1}(\gamma, \beta)\|\tilde{g}_\varepsilon - g\|_\infty\right)^p\right] > 0,$$

where $\inf_{\tilde{g}_\varepsilon}$ denotes the infimum over all estimators of g.

This theorem shows that the rate of convergence $\phi_\varepsilon(\gamma, \beta)$ cannot be improved on any estimators. We will next claim that for $0 < \gamma, \beta \leq 2$ there exist estimators attaining this rate. Before stating the corresponding result, we make some remarks on the properties of the rate $\phi_\varepsilon(\gamma, \beta)$.

Remark 1. There is an improvement due to the structure everywhere except for the trivial cases 1° and 2° discussed in Section 3. This corresponds to the zone $\{\mathcal{A} = (\gamma, \beta) : \beta > \gamma, \beta \geq 1\}$ which we refer to as *zone of improved rate* (cf. Figure 1). Indeed, when $\mathcal{A}$ belongs to this zone the effective smoothness is $\alpha_{\gamma,\beta} = \gamma$ (cf. (4)), and hence $\phi_\varepsilon(\gamma, \beta) = o(\psi_{\varepsilon,d}(\alpha_{\gamma,\beta}))$, as $\varepsilon \to 0$. Thus, the requirement (i) of Section 2 is met.

Remark 2. Parameter β can be viewed as a tuning parameter of the model: its choice can reduce the impact of the dimension d on the accuracy of estimation. Indeed, as the ratio d/β tends to 0, the rate $\phi_\varepsilon(\gamma, \beta)$ approaches either the one-dimensional Hölder rate $\psi_{\varepsilon,1}(\gamma)$ or the "almost parametric" rate $\varepsilon\sqrt{\ln(1/\varepsilon)}$. Thus, the requirement (iii) of Section 2 is met.

Remark 3. If $\beta \geq \gamma > 1$, $\beta < d(\gamma - 1) + 1$ the rate of convergence $\phi_\varepsilon(\gamma, \beta)$ does not depend on γ and coincides with the minimax rate $\psi_{\varepsilon,d}(\beta)$ associated to the d-dimensional Hölder class $\mathbb{H}_d(\beta, \cdot)$. This is rather surprising: in this zone the composite function $g = f \circ G$ can be estimated with the same rate as G, independently of how smooth is f. Such a behavior cannot be explained in terms of the smoothness because in the considered case the effective smoothness $\alpha_{\gamma,\beta}$ takes the value γ and not β (cf. (4)).

Remark 4. Theorem 1 obviously implies that the lower bound $\left(\varepsilon\sqrt{\ln(1/\varepsilon)}\right)^{\frac{2\gamma}{2\gamma+1+(d-1)/\beta}}$ is valid for all positive γ, β. Inspection of the proof shows that this bound is attained

at the least favorable functions that, for the particular case of $d = 2$, are of the form $f_0(\varphi_1(t_1) + \varphi_2(t_2))$ where f_0 is a function of Hölder smoothness γ and both functions φ_j are of Hölder smoothness β. So, for $d = 2$ the lower bound with rate $\left(\varepsilon\sqrt{\ln(1/\varepsilon)}\right)^{\frac{2\gamma}{2\gamma+1+1/\beta}}$ holds for such more restricted class of functions, whatever are γ and β. In particular, when $\gamma = \beta$, this lower rate becomes $\left(\varepsilon\sqrt{\ln(1/\varepsilon)}\right)^{\frac{2\beta^2}{2\beta^2+\beta+1}}$. Since $\frac{2\beta^2}{2\beta^2+\beta+1} < \frac{2\beta}{2\beta+1}$ this is always slower than the classical one-dimensional rate $\varepsilon^{\frac{2\beta}{2\beta+1}}$. On the other hand, a recent result of Horowitz and Mammen (7) shows that for $\gamma = \beta$ functions of the form $f_0(\varphi_1(t_1) + \varphi_2(t_2))$ can be estimated at the rate $\varepsilon^{\frac{2\beta}{2\beta+1}}$ in the L_2-norm. This phenomenon is very surprising because, in contrast to classical nonparametric estimation problems, we observe here a significant (and not only a logarithmic) deterioration of the rate when passing from the L_2-norm to the L_∞-norm.

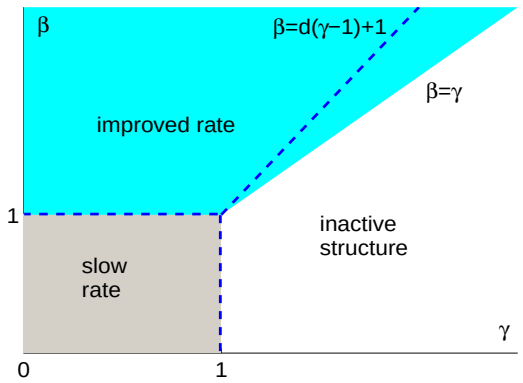

Figure 1. Zones of improved rate (cyan), of slow rate (grey) and of inactive structure (white). Dashed lines demarcate the zones of three different expressions for the exponent $\pi(\mathcal{A})$.

4.2 Outline of the estimation method

The exact definition of our estimator is given in Section 6. Here we only outline its construction. We suppose that $\mathcal{A} = (\gamma, \beta) \in (0, 2]^2$. The initial building block is a family of kernel estimators. In contrast to the classical kernel construction which involves a unique bandwidth parameter, the kernel $K_{\mathcal{J}}$ that we consider is determined by the triplet $\mathcal{J} = (\mathcal{A}, \vartheta, \lambda)$ where the *form* parameter $\mathcal{A}$ is the couple $(\gamma, \beta) \in (0, 2]^2$, the *orientation* parameter ϑ is a unit vector in $\mathbb{R}^d$ and λ is a positive real which we refer to as *size* parameter. We denote $\mathfrak{J}$ the set of all such triplets $\mathcal{J}$ and consider a family of kernel estimators $(\hat{g}_{\mathcal{J}}, \mathcal{J} \in \mathfrak{J})$ where for any $x \in [-1, 1]^d$ the estimator $\hat{g}_{\mathcal{J}}(x)$ of $g(x)$ is given by

$$\hat{g}_{\mathcal{J}}(x) \triangleq \int_{\mathcal{D}} K_{\mathcal{J}}(t - x) X_\varepsilon(dt).$$

We will see that, in general, the size parameter λ is not equivalent to the bandwidth of classical kernel estimator. In fact, the value of λ characterizes the bias of the estimator $\hat{g}_{\mathcal{J}}$ when the orientation of the window ϑ is locally "correct". Namely, the kernel $K_{\mathcal{J}}$ is chosen in such a way that for each $x \in [-1, 1]^d$ the bias of $\hat{g}_{\mathcal{J}}$ is of the order $O(\lambda)$ if $\vartheta = \vartheta_0^x$ is collinear to the gradient $\nabla G(x)$.

The estimation method proceeds in three steps, and the basic device underlying the construction of the optimal estimation method is the notion of the *local model*. It is an important feature of the composition structure that different local models arise in different subsets of the zone of improved rate.

Step 1. Specifying a collection of local models. The underlying function g with a complicated global structure can have a simple local structure. However, the local structure depends on the function itself. Therefore, g can be only described by a *collection of local models*. In our case, this collection is indexed by a finite-dimensional parameter which can be considered as a nuisance parameter. Specifically, we pass from the global composition model defined in Section 3 to a family of local models $\{\mathcal{M}_{\mathcal{J}}(x),\, \mathcal{J} \in \mathfrak{J},\, x \in [-1,1]^d\}$ where the type of each local model $\mathcal{M}_{\mathcal{J}}(x)$, $\mathcal{J} = (\mathcal{A}, \vartheta, \lambda)$, is determined by $\mathcal{A}$, while ϑ and λ are the local orientation and size parameters. We will show that, depending on the value of $\mathcal{A} = (\gamma, \beta)$ (cf. Figure 2), our global model induces only two types of local models: a *local single index model* and a *combined local model*. The latter combines elements of both single index and additive models. This responds to the requirement (ii) of Section 2.

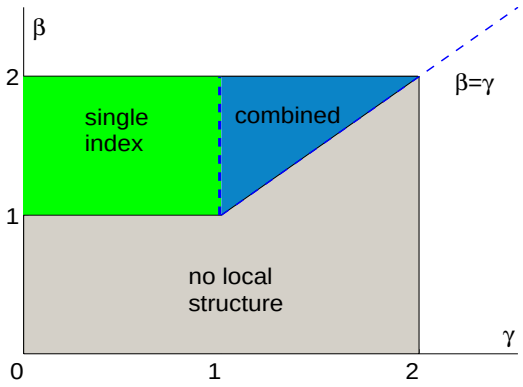

Figure 2. Types of local structures.

1°. *Local single index model:* $\gamma \le 1,\ 1 < \beta \le 2$.

In this domain of γ, β, using the smoothness properties of functions f and G it is not hard to show that in the ball $B_{\lambda,x}(\mathcal{A}) = \left\{t \in \mathbb{R}^d :\ \ \|t - x\| \le \lambda^{\frac{1}{\gamma\beta}}\right\}$ the composite function $g(\cdot)$ can be approximated with the accuracy $O(\lambda)$ by the function $f\big(G(x) + \vartheta^T[\cdot - x]\big)$. Here $\vartheta = \vartheta_0^x$ is a unit vector collinear to the gradient $\nabla G(x)$. Indeed, since the inner function G belongs to $H_d(\beta, L_2)$, for any $x, y \in \mathcal{D}$ we have

$$G(t) = G(x) + \nabla G(x)^T(t - x) + B_x(t), \ \text{ with } B_x(t) \le L_2\|t - x\|^\beta. \tag{8}$$

Next, using the fact that $f \in \mathbb{H}_1(\gamma, L_1)$, we conclude that $g(t) = f(G(t))$ admits the representation

$$g(t) = Q_x(t) + C_x(t),$$

where

$$Q_x(t) = f(G(x) + \nabla G(x)^T(t - x)) \ \text{ and } \ |C_x(t)| \le L_1|B_x(t)|^\gamma \le L_1 L_2^\gamma \|t - x\|^{\gamma\beta}.$$

In other words, for any kernel K with the support on the ball $B_\lambda(\mathcal{A}) = \left\{t \in \mathbb{R}^d :\ \|t\| \le \lambda^{1/\gamma\beta}\right\}$ and such that $\int K(y)dy = 1$,

$$\int K(t - x)[g(t) - Q_x(t)]dt = O(\lambda). \tag{9}$$

We understand the relation (9) as the definition of the local single index model Q_x of g. The choice of the approximation kernel for the function g is naturally suggested by the form of the local model Q_x together with the bound (9): the kernel $K_{\mathcal{J}}$ can be taken as the indicator function of a hyperrectangle normalized by its volume and oriented in such a way that $\nabla G(x)$ is collinear to the first basis vector in $\mathbb{R}^d$. The sides of the hyperrectangle are chosen to have the lengths $l_1 = \lambda^{\frac{1}{\gamma}}$ and $l_j = \lambda^{\frac{1}{\gamma\beta}}, j = 1, \dots, d-1$.

2°. *Combined local model:* $1 < \gamma \le \beta \le 2$.

Let M_ϑ be an orthogonal matrix with the first column equal to $\vartheta = \vartheta_0^x$, and let $y = M_\vartheta^T(t-x),\ t \in \mathbb{R}^d$. We denote with y_j the jth component of y and consider the set

$$\mathcal{X}_{\lambda,x}(\mathcal{A}) = \left\{ t \in \mathbb{R}^d :\ |y_1| \le \lambda^{\frac{1}{\beta}},\ \|y\| \le \lambda^{\frac{1}{\gamma\beta}},\ |y_1|^{\gamma-1}\|y\|^\beta \le \lambda \right\}. \tag{10}$$

We show that the estimation of the composite function g at x can be reduced to the problem of estimation under the local model

$$Q_x(y) = q_x(y_1) + P_x(y_2, \dots, y_d),$$

where $q_x \in \mathbb{H}_1(\gamma, L_1 L_2^\gamma)$ and $P_x \in \mathbb{H}_{d-1}(\beta, 2L_1L_2)$ on the set $\mathcal{X}_{\lambda,x}(\mathcal{A})$. This local model is established in an unknown coordinate system determined by the parameter $\vartheta = \vartheta_0^x$. Since y_1 is the coordinate of the projection on ϑ_0^x, the component $q_x(y_1)$ constitutes an element of single index structure. An element of additive structure comes from the separation of Q_x into the sum of two functions depending on non-intersecting sets of coordinates.

The explanation of the local model represented by Q_x on the set $\mathcal{X}_{\lambda,x}(\mathcal{A})$ is provided by the following argument. Using the smoothness properties of functions f and G, we obtain due to the inclusions $f \in \mathbb{H}_1(\gamma, L_1)$, $G \in \mathbb{H}_d(\beta, L_2)$:

$$\begin{aligned} g(t) &= f(G(x) + \nabla G(x)^T(t-x)) + f'(G(x) + \nabla G(x)^T(t-x))B_x(t) + C_x(t) \\ &= f(G(x) + \nabla G(x)^T(t-x)) + f'(G(x))B_x(t) + D_x(t) + C_x(t), \end{aligned}$$

where

$$\begin{aligned} |C_x(t)| &\le C(L_1, L_2, \gamma)\|t-x\|^{\gamma\beta}, \\ |D(x,y)| &\le C(L_1, L_2)\frac{|\nabla G(x)^T(t-x)|}{\|\nabla G(x)\|}\|t-x\|^\beta, \end{aligned}$$

and the function $B_x(t)$, which is defined in (8), belongs to the class $\mathbb{H}_d(\beta, 2L_2)$. In the transformed coordinates (determined by the orthogonal matrix M_ϑ) we may write

$$g(t) = g(x + M_\vartheta y) = q(y_1) + \tilde{B}_x(y) + \tilde{D}_x(y) + \tilde{C}_x(y), \tag{11}$$

where

$$|\tilde{D}_x(y) + \tilde{C}_x(y)| \le C(L_1, L_2, \gamma)(|y_1|^{\gamma-1}\|y\|^\beta + \|y\|^{\gamma\beta}). \tag{12}$$

and $\tilde{B}_x \in \mathbb{H}_d(\beta, 2L_2)$. The latter inclusion leads to

$$\left| \tilde{B}_x(y) - P_x(y_2, \dots, y_d) - y_1 \frac{\partial}{\partial y_1}\tilde{B}_x(0, y_2, \dots, y_d) \right| \le 2L_2|y_1|^\beta, \tag{13}$$

where $P_x(y_2, \dots, y_d) = \tilde{B}_x(0, y_2, \dots, y_d)$. Let again K be a kernel such that $\int K(t)dt = 1$, supported on $\mathcal{X}_{\lambda,x}(\mathcal{A})$. Then

$$\int K(y-x)[g(x + M_\vartheta y) - Q_x(y)]dy = O(\lambda) \tag{14}$$

if K is *symmetric in* y_1. We understand this property as the definition of the combined local model Q_x for the composite function g.

We conclude that if $\mathcal{A}$ belongs to the zone marked as "combined" in Figure 2, the *global* structural assumption that the underlying function is a composite one leads automatically to a *local* structure containing elements of both single index and additive models.

A good kernel $K_{\mathcal{J}}$ for the zone of combined local model should be supported on the right window $\mathcal{X}_{\lambda,x}(\mathcal{A})$, possess small bias on both single-index component q_x and "regular" component P_x and have a small L_2-norm to ensure small variance of the stochastic term of the estimation error. Construction of such a kernel is a rather involved task (cf. Section 7.2). Using a rectangular kernel, as for the local single-index model, does not give a solution, since it leads to suboptimal estimation rates.

As we see, the definition of local model has two ingredients: the neighborhood (window) and the local structure within the window. For the local single index model the window is just an Euclidean ball, whereas for the combined local model the window is the set $\mathcal{X}_{\lambda,x}(\mathcal{A})$ which has quite a nonstandard form (cf. Figure 3).

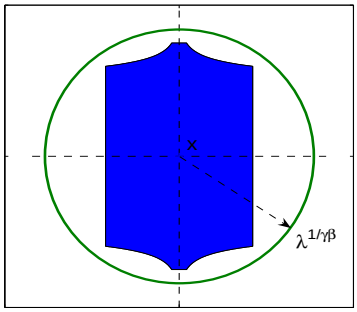

Figure 3. Window for the combined local model, $d = 2$.

Step 2.Optimizing the size parameter and specifying candidate estimators. Once the local model is determined and the corresponding kernel is constructed we can chose the size parameter $\lambda = \lambda_\varepsilon(\mathcal{A})$ in an optimal way. To do it we optimize our sup-norm risk with respect to λ, i.e., we get the value λ which realizes the balance of bias and variance terms of the risk in the ideal case where the orientation $\vartheta = \vartheta_0^x$ is "correct" for all x.

Recall that the kernel $K_{\mathcal{J}}$ supported on the window is chosen in such a way that the bias of the kernel estimator $\hat{g}_{\mathcal{J}}$, for the "correct" orientation ϑ, is of the order $O(\lambda)$ on every local model. Thus, the bias-variance balance relation for the sup-norm loss can be written in the form

$$\lambda \asymp \varepsilon\sqrt{\ln 1/\varepsilon}\|K_{\mathcal{J}}\|_2. \tag{15}$$

We will see that $\|K_{\mathcal{J}}\|_2$ depends on $\mathcal{A}$ and λ but does not depend on ϑ. This will allow us to choose the optimal value $\lambda_\varepsilon(\mathcal{A})$ independent of ϑ. For instance, for the local single index model the kernel $K_{\mathcal{J}}$ is just a properly scaled and rotated indicator of a hyperrectangle. In this particular case the bias-variance balance (15) can be written in the form

$$\lambda \asymp \frac{\varepsilon\sqrt{\ln 1/\varepsilon}}{\sqrt{\text{volume of hyperrectangle}}} = \varepsilon\left(\frac{\ln 1/\varepsilon}{\lambda^{\frac{1}{\gamma}+\frac{d-1}{\gamma\beta}}}\right)^{1/2}.$$

Note that in this case $\lambda_\varepsilon(\mathcal{A}) \asymp \phi_\varepsilon(\gamma,\beta)$, where $\phi_\varepsilon(\gamma,\beta)$ is defined in (7). On the other hand, to guarantee that the same relation $\lambda_\varepsilon(\mathcal{A}) \asymp \phi_\varepsilon(\gamma,\beta)$ holds in the zone of combined local model need a rather sophisticated construction of the kernel $K_{\mathcal{J}}$ (cf. Section 7.2).

With $\lambda_\varepsilon(\mathcal{A})$ being chosen, we obtain a family of kernel estimators

$$\left\{\hat{g}_{\mathcal{J}}(x),\ \ \mathcal{J}=(\mathcal{A},\vartheta,\lambda_\varepsilon(\mathcal{A}))\in\mathfrak{J},\ \ x\in[-1,1]^d\right\}. \tag{16}$$

For a fixed $x\in[-1,1]^d$ this family only depends on two parameters, $\mathcal{A}$ and ϑ.

Step 3. Aggregating the estimators. We now choose an estimator from the family (16) which corresponds to some $\hat{\mathcal{J}}\in\mathfrak{J}$ selected in a data-dependent way, and define our final estimator as a piecewise-constant approximation of the function $x\mapsto\hat{g}_{\hat{\mathcal{J}}}(x)$. To choose $\hat{\mathcal{J}}$ we apply an aggregation procedure which is a special case of the method of aggregation of linear estimators proposed in (17).

We introduce a discrete grid on the unit sphere $\{\vartheta\in\mathbb{R}^d:\ \|\vartheta\|=1\}$, and we divide the domain of definition of x into small blocks. For each block, we consider a finite set of estimators $\hat{g}_{\mathcal{J}}(x)$ extracted from the family (16), with x which is fixed as the center x_0 of the block and all the ϑ on the grid. We then select a data-dependent $\hat{\vartheta}$ in the grid applying our aggregation procedure to this finite set. The value of our final estimator $g^*_{\mathcal{A},\varepsilon}$ on this block is constant and is defined as $g^*_{\mathcal{A},\varepsilon}(x)\equiv\hat{g}_{(\mathcal{A},\hat{\vartheta},\lambda_\varepsilon(\mathcal{A}))}(x_0)$. We thus get a piecewise-constant estimator $g^*_{\mathcal{A},\varepsilon}$ on $[-1,1]^d$ which depends only on $\mathcal{A}$ and on the observations (the exact definition of $g^*_{\mathcal{A},\varepsilon}$ is given in Section 6).

Remark 5. If $\mathcal{A}$ is unknown we need simultaneous adaptation to $\mathcal{A}$ and to ϑ, i.e., to the smoothness and to the local structure of the underlying function. Note, however, that parameters $\mathcal{A}$ and ϑ are not independent. In particular, $\mathcal{A}$ determines the form of the neighborhood where we have an *unknown* local structure depending on ϑ. This is important because our construction of the family of estimators $\{\hat{g}_{\mathcal{J}},\mathcal{J}\in\mathfrak{J}\}$ heavily relates on the local representation of the model. For example, if the family $\{\hat{g}_{\mathcal{J}},\mathcal{J}\in\mathfrak{J}\}$ does not contain an estimator corresponding to the correct local structure, the choice from this family cannot even guarantee consistency. Another difficulty is that different values of $\mathcal{A}$ can correspond to different *types* of local models (cf. Figure 2). Therefore, if $\mathcal{A}$ is totally unknown (fully adaptive estimation) then both the type of local structure and the form of the corresponding window are unknown. So, we see that adaptive estimation of composite functions is more difficult than classical adaptation to the unknown smoothness as considered, for example, in (14–16). In a forthcoming paper we will show that it is possible to adapt to unknown *type* of the local structure (including adaptation to the local orientation ϑ) under certain known restrictions on $\mathcal{A}$. We will call this partial adaptation. Partial adaptive procedure can be constructed in a similar way as discussed above. A difference is that we need to introduce a grid not only on the values of ϑ, but also on those of $\mathcal{A}$, and we aggregate estimators corresponding to the product of both grids.

4.3 Upper bounds on the risk of the estimators

Define the following three domains of values of $\mathcal{A}=(\gamma,\beta)$ contained in $(0,2]^2$ (cf. Figure 4).

$$\begin{aligned}\mathcal{P}_1 &= \left\{\mathcal{A}:\gamma\le 1,\ 1<\beta\le 2\right\},\\ \mathcal{P}_2 &= \left\{\mathcal{A}:1<\gamma\le\beta\le 2,\ \beta\ge d(\gamma-1)+1\right\},\\ \mathcal{P}_3 &= \left\{\mathcal{A}:1<\gamma\le\beta\le 2,\ \beta<d(\gamma-1)+1\right\}.\end{aligned} \tag{17}$$

In view of the above discussion, these are exactly the zones where improved rates occur and where the local structure is active. For the sake of completeness, we consider also the

remainder zone (zone of no local structure):

$$\mathcal{P}_4 = (0,1]^2 \cup \{(\gamma,\beta) : 1 \le \beta < \gamma \le 2\}.$$

As we will see it later, the optimal kernels $K_{\mathcal{J}}$ are defined differently for each of these zones.

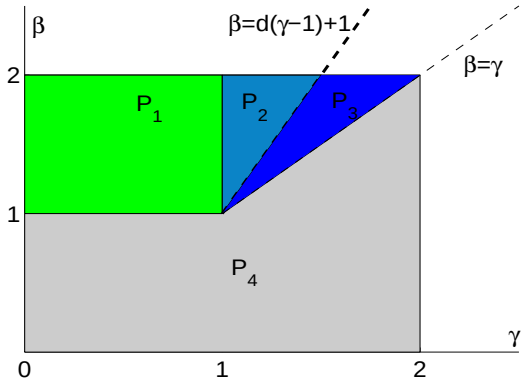

Figure 4. Classification of zones within $(0,2]^2$.

Theorem 2. *Let $\phi_\varepsilon(\gamma,\beta)$ be as in (7). For any $\mathcal{A} = (\gamma,\beta) \in (0,2]^2 \setminus \mathcal{P}_2$ and any $p > 0$ the estimator $g^*_{\mathcal{A},\varepsilon}$ satisfies*

$$\limsup_{\varepsilon\to 0} \sup_{g\in\mathbb{H}(\mathcal{A},\mathcal{L})} \mathbb{E}_g \left[\left(\phi_\varepsilon^{-1}(\gamma,\beta)\|g^*_{\mathcal{A},\varepsilon} - g\|_\infty\right)^p\right] < \infty.$$

For any $\mathcal{A} = (\gamma,\beta) \in \mathcal{P}_2$ and any $p > 0$ the estimator $g^*_{\mathcal{A},\varepsilon}$ satisfies

$$\limsup_{\varepsilon\to 0} \sup_{g\in\mathbb{H}(\mathcal{A},\mathcal{L})} \mathbb{E}_g \left[\left(\left[\ln\ln(1/\varepsilon)\right]^{-1}\phi_\varepsilon^{-1}(\gamma,\beta)\|g^*_{\mathcal{A},\varepsilon} - g\|_\infty\right)^p\right] < \infty.$$

Combining Theorems 1 and 2 we conclude that $\phi_\varepsilon(\gamma,\beta)$ is the minimax rate of convergence for the class $\mathbb{H}(\mathcal{A},\mathcal{L})$ if $\mathcal{A} = (\gamma,\beta) \in (0,2]^2 \setminus \mathcal{P}_2$, and that it is near minimax (up to the $\ln\ln(1/\varepsilon)$ factor) if $\mathcal{A} = (\gamma,\beta) \in \mathcal{P}_2$. Therefore, our estimator $g^*_{\mathcal{A},\varepsilon}$ is respectively rate optimal or near rate optimal on $\mathbb{H}(\mathcal{A},\mathcal{L})$.

Theorem 2 can be viewed as a result on adaptation to the unknown local structure of the function to be estimated: the estimator $g^*_{\mathcal{A},\varepsilon}$ locally adapts to the "correct" orientation ϑ which is a vector collinear to the gradient $\nabla G(x)$ in a neighborhood of x.

5 Extensions

5.1 Related statistical models

1. We consider here the Gaussian white noise model because its analysis requires a minimum of technicalities. Composition structures can be studied for more realistic models, such as nonparametric regression with random design, nonparametric density estimation and classification. Note that our theorems can be directly transposed to gaussian nonparametric regression model with fixed equidistant design using the equivalence of experiments argument (cf. (4; 21)). Note also that results similar to ours have been recently obtained for the problem of testing hypotheses about composite functions in the Gaussian white noise model (18).

2. We restrict our study to the sup-norm loss and to the Hölder smoothness classes. A natural extension would be to consider models where the risk is described by other norms and other smoothness classes. Typical candidates here are Sobolev and Besov classes, the classes of monotone or convex functions. The case of functional classes with anisotropic smoothness is of interest as well. Estimation in other norms may lead to unexpected effects (cf. Remark 4).

3. We consider only the simplest composition $f(G)$, where $f : \mathbb{R} \to \mathbb{R}$ and $G : \mathbb{R}^d \to \mathbb{R}$. A more general description could be $f(G_1, \dots, G_k)$, where $f : \mathbb{R}^k \to \mathbb{R}$ and $G_s : \mathbb{R}^{d_s} \to \mathbb{R}$, $s = 1, \dots, k$, and $d_1 + \cdots + d_k = d$.

4. A related more complex modeling can be based on Kolmogorov's theorem of representation of a continuous function of several variables by compositions and sums of functions of one variable addition (13; 22).

5.2 Possible refinements

1. In this paper we treat only the case $\mathcal{A} \in (0, 2]^2$. Extension to $\mathcal{A} \notin (0, 2]^2$ remains an open problem. However, our lower bound (Theorem 1) is valid for all $\mathcal{A} \in \mathbb{R}^2_+$. We believe that it cannot be improved. This conjecture is supported by recent results on a hypothesis testing problem with composite functions (18) which is closely related to our estimation problem. The upper bound proved in (18) for all $\mathcal{A} \in \mathbb{R}^2_+$ in the problem of testing hypotheses coincides with our lower bound.

2. The rate of convergence of the minimax procedure (cf. Theorem 2) in the zone $\mathcal{P}_2$ contains an additional $\ln \ln(1/\varepsilon)$ factor, as compared to the lower bound of Theorem 1. This deterioration of the rate is due to the method of aggregation of linear estimators that we use and does not seem to be unavoidable.

3. In Remark 5 we discussed a possible construction for partial adaptation. The ultimate goal of adaptation is, however, to find an estimator which is *totally parameter free*. Such an estimator should achieve the minimax rate (6) simultaneously for all $\mathcal{A} \in \mathbb{R}^2_+$.

6 Definition of the estimator

We first introduce some notation. For a bounded function $K \in L_1(\mathbb{R}^d)$ and $p \geq 1$ we denote by $\|K\|_p$ its L_p-norm and by $K * g$ its convolution with a bounded function g:

$$\|K\|_p = \left(\int |K(t)|^p dt \right)^{1/p}, \quad [K * g](x) = \int K(t - x) g(t) dt, \quad x \in \mathbb{R}^d$$

(here and in the sequel $\int = \int_{\mathbb{R}^d}$). We denote $\mathcal{J} \triangleq (\mathcal{A}, \vartheta, \lambda)$ where $\mathcal{A} = (\gamma, \beta) \in (0, 2]^2$, ϑ is a unit vector in $\mathbb{R}^d$ and $\lambda > 0$. The class of all such triplets $\mathcal{J}$ is denoted by $\mathfrak{J}$.

Given a unit vector ϑ, let $M_\vartheta \in \mathbb{R}^{d \times d}$ stand for an orthogonal matrix with the first column equal to ϑ. The collection of the kernels we consider in the sequel is defined as

$$K_{\mathcal{J}}(x) = \mathsf{K}_{(\mathcal{A},\lambda)}(M_\vartheta^T x)$$

where $\mathsf{K}_{(\mathcal{A},\lambda)} : \mathbb{R}^d \to \mathbb{R}$ is a kernel that will be defined in Section 7. Next, for any $\mathcal{J}', \mathcal{J}, \in \mathfrak{J}$ and all $t \in \mathbb{R}^d$ we define the *convoluted kernel*

$$K_{\mathcal{J}\prime * \mathcal{J}}(t) = \int K_{\mathcal{J}\prime}(t-y) K_{\mathcal{J}}(y) dy$$

and the difference

$$\Delta_{\mathcal{J}\prime} K_{\mathcal{J}\prime * \mathcal{J}} = K_{\mathcal{J}\prime * \mathcal{J}} - K_{\mathcal{J}\prime}.$$

Note that, by definition, the kernel $K_{\mathcal{J}}$ is symmetric, i.e., $K_{\mathcal{J}}(t) = K_{\mathcal{J}}(-t)$, and

$$K_{\mathcal{J}\prime * \mathcal{J}} = K_{\mathcal{J} * \mathcal{J}\prime}. \tag{18}$$

For all $\mathcal{J} \in \mathfrak{J}$ and all $x \in [-1,1]^d$ set

$$\hat{g}_{\mathcal{J}}(x) = \int_{\mathcal{D}} K_{\mathcal{J}}(t-x) X_\varepsilon(dt),$$

and for all $\mathcal{J}', \mathcal{J} \in \mathfrak{J}$ define the *convoluted estimator*

$$\hat{g}_{\mathcal{J}' * \mathcal{J}}(x) = \int_{\mathcal{D}} K_{\mathcal{J}' * \mathcal{J}}(t-x) X_\varepsilon(dt).$$

In what follows we assume that ε is small enough so that $\ln\ln(1/\varepsilon) > 0$ and that in the above expressions and in all the subsequent expressions containing convolutions with the kernels we can replace $\int_{\mathcal{D}}$ by $\int_{\mathbb{R}^d}$ (this is possible for small ε since all the kernels that we consider are compactly supported). We also define

$$\Delta_{\mathcal{J}'} \hat{g}_{\mathcal{J}' * \mathcal{J}}(x) = \hat{g}_{\mathcal{J}' * \mathcal{J}}(x) - \hat{g}_{\mathcal{J}'}(x)$$

and set

$$\mathbf{TH}_\varepsilon(\mathcal{J}', \mathcal{J}) = C(p,d) \left(\|K_{\mathcal{J}'}\|_1 + \|K_{\mathcal{J}}\|_1 \right) \|K_{\mathcal{J}'}\|_2 \, \varepsilon \sqrt{\ln(1/\varepsilon)},$$

where $C(p,d) = 2 + \sqrt{4p + 8d}$.

To define the estimator we first introduce a discrete grid on the set of indices $\mathfrak{J}$. We discretize only the ϑ-coordinate of $\mathcal{J}$. Recall that ϑ takes values on the Euclidean unit sphere $\mathbb{S}$ in $\mathbb{R}^d$.

Discretization Let $\mathbb{S}_\varepsilon \subset \mathbb{S}$ be an ε^2-net on $\mathbb{S}$, i.e., a finite set such that

$$\forall \vartheta \in \mathbb{S} \; \exists \vartheta' \in \mathbb{S}_\varepsilon : \; \|\vartheta - \vartheta'\| \le \varepsilon^2,$$

and $\mathrm{card}(\mathbb{S}_\varepsilon) \le (\sqrt{d}\varepsilon^{-2})^d$. W.l.o.g. we will assume that $(1, 0, \ldots, 0) \in \mathbb{S}_\varepsilon$.

Fix $\mathcal{A} \in (0,2]^2$ and define $\lambda_\varepsilon(\mathcal{A})$ as a solution in λ of the bias-variance balance equation

$$c_{11}\lambda = \varepsilon \sqrt{\ln(1/\varepsilon)} \|\mathsf{K}_{(\mathcal{A},\lambda)}\|_2 \tag{19}$$

where c_{11} is a constant in Proposition 1 below, depending only on $\mathcal{A}$, $\mathcal{L}$ and d. Finally we define the following grid on the values of $\mathcal{J}$:

$$\mathfrak{J}_{\mathrm{grid}} \triangleq \left\{ \mathcal{J} = (\mathcal{A}, \vartheta, \lambda_\varepsilon(\mathcal{A})) : \; \vartheta \in \mathbb{S}_\varepsilon \right\} \subset \mathfrak{J}.$$

Acceptability For a given $x \in [-1,1]^d$ we define a subset $\hat{\mathfrak{T}}_x$ of $\mathfrak{J}_{\rm grid}$ as follows:

$$\mathcal{J} \in \hat{\mathfrak{T}}_x \Longleftrightarrow \left|\Delta_{\mathcal{J}'}\hat{g}_{\mathcal{J}'*\mathcal{J}}(x)\right| \le \mathbf{TH}_\varepsilon(\mathcal{J}',\mathcal{J}), \quad \forall \mathcal{J}' \in \mathfrak{J}_{\rm grid}.$$

Any value $\mathcal{J}$ belonging to $\hat{\mathfrak{T}}_x$ is called *acceptable*.

Note that the threshold $\mathbf{TH}_\varepsilon(\mathcal{J}',\mathcal{J})$ can be bounded from above and replaced in all the definitions by a value that does not depend on $\mathcal{J},\mathcal{J}' \in \mathfrak{J}_{\rm grid}$. In fact, either $\mathbf{TH}_\varepsilon(\mathcal{J}',\mathcal{J}) \asymp \lambda_\varepsilon(\mathcal{A})$ if $\mathcal{A} \in \mathcal{P}_1 \cup \mathcal{P}_3$ or $\mathbf{TH}_\varepsilon(\mathcal{J}',\mathcal{J}) \asymp \ln\ln(1/\varepsilon)\lambda_\varepsilon(\mathcal{A})$ if $\mathcal{A} \in \mathcal{P}_2$.

Estimation at a fixed point For any $x \in [-1,1]^d$ such that $\hat{\mathfrak{T}}_x \neq \emptyset$ we select an arbitrary $\hat{\mathcal{J}}_x$ from the set $\hat{\mathfrak{T}}_x$. Note that the set $\hat{\mathfrak{T}}_x$ is finite, so a measurable choice of $\hat{\mathcal{J}}_x$ is always possible; we assume that such a choice is effectively done. We then define the estimator $g^{**}(x)$ as follows:

$$g^{**}(x) \triangleq \begin{cases} \hat{g}_{\hat{\mathcal{J}}_x}(x) & \text{if} \quad \hat{\mathfrak{T}}_x \neq \emptyset, \\ 0 & \text{if} \quad \hat{\mathfrak{T}}_x = \emptyset. \end{cases} \tag{20}$$

Global estimator The estimator g^{**} is defined for all $x \in [-1,1]^d$ and we could consider $x \mapsto g^{**}(x),\ x \in [-1,1]^d$, as an estimator of the function g. However, the measurability of this mapping is not a straightforward issue. To skip the analysis of measurability, we use again a discretization. Introduce the following cubes in $\mathbb{R}^d$:

$$\Pi_\varepsilon(z) = \bigotimes_{k=1}^d \left[\varepsilon^2(z_k-1), \varepsilon^2 z_k\right],\ z = (z_1,\dots,z_d) \in \mathbb{Z}^d.$$

For any $x \in [-1,1]^d$ we consider $z(x) \in \mathbb{Z}^d$ such that x belongs to the cube $\Pi_\varepsilon(z(x))$, and a piecewise constant estimator $g^{**}(z(x))$. Our final estimator is a truncated version of $g^{**}(z(x))$:

$$g^*_{\mathcal{A},\varepsilon}(x) \triangleq \begin{cases} g^{**}(z(x)) & \text{if} \quad |g^{**}(z(x))| \le \ln\ln(1/\varepsilon), \\ \ln\ln(1/\varepsilon)\,\mathrm{sign}(g^{**}(z(x))) & \text{if} \quad |g^{**}(z(x))| > \ln\ln(1/\varepsilon). \end{cases} \tag{21}$$

Thus, the resulting procedure $g^*_{\mathcal{A},\varepsilon}$ is piecewise constant on the cubes $\Pi_\varepsilon(z) \subset [-1,1]^d,\ z \in \mathbb{Z}^d$.

7 Construction of the kernel

In this section, as well as in the Appendix, we will distinguish between the couple of "true" parameters $\mathcal{A}_0 = (\gamma,\beta)$ and a variable couple of parameters $\mathcal{A} = (a,b) \in (0,2]^2$. This is done to state the lemmas in a form convenient to be applied in the context of adaptation to unknown (γ,β) which will be treated in our forthcoming work.

Depending on the value of $\mathcal{A}$ we use different constructions of $\mathsf{K}_{(\mathcal{A},\lambda)}$. Our objective is to obtain $K_\mathcal{J}$ with suitable approximation properties for each $\mathcal{J} \in \mathfrak{J}$. Let us summarize here the main requirements on the kernel:

1. Convolution of the kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ with the "local model" of g corresponding to $\mathcal{A}$ should approximate g with the accuracy $O(\lambda)$. Furthermore, the kernel should be localized, i.e., it should vanish outside of the window where the local structure is valid.

2. A basic characteristic of the kernel is its L_2-norm which determines the variance of the kernel estimator. Our objective is to achieve its minimal value.

3. As we will see it later, the L_1-norm of the kernel is also an important parameter of the proposed estimation procedure. Our objective will be to keep the L_1-norm as small as possible.

We use different kernels $\mathsf{K}_{(\mathcal{A},\lambda)}$ for $\mathcal{A}$ belonging to different zones $\mathcal{P}_i$ (cf. Figure 4). The construction of $\mathsf{K}_{(\mathcal{A},\lambda)}$ is trivial when $\mathcal{A}$ is in the zone $\mathcal{P}_4$ of no local structure. In this case a basic boxcar kernel tuned to the effective smoothness of the composite function can be used. Observe that when $\mathcal{A} \in (0,1]^2$ the effective smoothness of the composite function equals to ab, and when $\mathcal{A} = (a,b)$ satisfies $1 < b \le a \le 2$ the effective smoothness is b. So, we define the kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ for the zone $\mathcal{P}_4$ as follows:

$$\mathsf{K}_{(\mathcal{A},\lambda)}(y) = \begin{cases} \left(2\lambda^{\frac{1}{ab}}\right)^{-d} \mathbb{I}_{\left[-\lambda^{\frac{1}{ab}},\lambda^{\frac{1}{ab}}\right]^d}(y) & \text{if} \quad \mathcal{A} = (a,b) \in (0,1]^2, \\ \left(2\lambda^{1/b}\right)^{-d} \mathbb{I}_{[-\lambda^{1/b},\lambda^{1/b}]^d}(y) & \text{if} \quad 1 < b < a \le 2. \end{cases}$$

Here $\mathbb{I}_A(\cdot)$ stands for the indicator function of a set A. The following lemma is straightforward.

Lemma 1. *For any $\mathcal{A}_0 = (a,b) \in \mathcal{P}_4$, $\lambda > 0$ and $x \in [-1,1]^d$, we have*

$$\sup_{g \in \mathbb{H}(\mathcal{A},\mathcal{L})} |[\mathsf{K}_{(\mathcal{A},\lambda)} * g](x) - g(x)| \le c_0 \lambda,$$

where the constant c_0 depends only on $\mathcal{L}$ and d. Furthermore,

$$\|\mathsf{K}_{(\mathcal{A},\lambda)}\|_1 = 1 \quad \textit{and} \quad \|\mathsf{K}_{(\mathcal{A},\lambda)}\|_2 = \begin{cases} \left(2\lambda^{\frac{1}{ab}}\right)^{-d/2}, & (a,b) \in (0,1]^2 \\ \left(2\lambda^{\frac{1}{b}}\right)^{-d/2}, & 1 < b < a \le 2. \end{cases}$$

We turn now to the analysis of cases with active local structure. We start with the zone $\mathcal{P}_1$ of local single model.

7.1 Kernel for the local single index model

The zone of local single-index model is $\mathcal{P}_1 = \big\{\mathcal{A} = (a,b) : a \le 1,\ 1 < b \le 2\big\}$. For any $\mathcal{A} \in \mathcal{P}_1$ and $\lambda > 0$ consider the hyperrectangle

$$\Pi_\lambda(\mathcal{A}) = \left[-\lambda^{1/a}, \lambda^{1/a}\right] \times \left[-\lambda^{\frac{1}{ab}}, \lambda^{\frac{1}{ab}}\right]^{d-1}$$

and define the kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ as follows:

$$\mathsf{K}_{(\mathcal{A},\lambda)} = \left(2^d \lambda^{\frac{1}{a}+\frac{d-1}{ab}}\right)^{-1} \mathbb{I}_{\Pi_\lambda(\mathcal{A})}(y), \quad y \in \mathbb{R}^d. \tag{22}$$

Approximation property of the kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ Let $q : \mathbb{R} \to \mathbb{R}$ and $B : \mathbb{R}^d \to \mathbb{R}$ be functions such that, for given $a \in (0,1]$,

$|q(x) - q(y)| \le L|x-y|^a, \quad \forall\, x, y \in \mathbb{R}^d,$

$\sup_{x \in \mathbb{R}^d} |B(x)| \le c_1$

where $c_1 > 0$, $L > 0$ are constants. We denote by $\mathfrak{A}(a)$ the set of all pairs of functions (q, B) satisfying these restrictions. Define

$$Q(y) = q(y_1) + B(y)\|y\|^{ab}, \quad \forall\ y \in \mathbb{R}^d.$$

We have the following evident result:

Lemma 2. *For any $\mathcal{A} = (a, b) \in \mathcal{P}_1$ and $\lambda > 0$ we have*

$$(i) \quad \sup_{(q,B)\in\mathfrak{A}(a)} \left|\left[\mathsf{K}_{(\mathcal{A},\lambda)} * Q\right](0) - q(0)\right| \le c_2\lambda$$

where c_2 is a constant depending only on L, c_1 and d. Moreover,

$$(ii) \quad \|\mathsf{K}_{(\mathcal{A},\lambda)}\|_1 = 1 \quad \textit{and} \quad \|\mathsf{K}_{(\mathcal{A},\lambda)}\|_2 = \left(2^d \lambda^{\frac{1}{a}+\frac{d-1}{ab}}\right)^{-1/2}.$$

7.2 Kernels for the combined local model

The zone of combined local model is $\mathcal{P}_2 \cup \mathcal{P}_3 = \{\mathcal{A} = (a, b) : 1 < a \le b \le 2\}$. The definition of the kernel in this case is more involved. Indeed, taking $\mathsf{K}_{(\mathcal{A},\lambda)}$ as a simple product of boxcar kernels (22) results for $\mathcal{A} \in \mathcal{P}_2 \cup \mathcal{P}_3$ in too large approximation error.

Our aim is to construct a smoothing kernel $\mathsf{K}_{(\mathcal{A},\lambda)} : \mathbb{R}^d \to \mathbb{R}$ with the following properties:

– for some $c > 0$, it should vanish outside the set (cf. (10))

$$\left\{y \in \mathbb{R}^d :\ |y_1| \le c\lambda^{\frac{1}{b}},\ \|y\| \le c\lambda^{\frac{1}{ab}},\ |y_1|^{a-1}\|y\|^b \le c\lambda\right\}.$$

– for a function $q(y_1)$ of the first component y_1 of $y \in \mathbb{R}^d$, the "characteristic size" of $\mathsf{K}_{(\mathcal{A},\lambda)}$ should be $\lambda^{\frac{1}{a}}$; for a function $Q(y_2, ..., y_d)$ of the remaining components $y_2, ..., y_d$ it should be $\lambda^{\frac{1}{b}}$. Namely, we want to ensure the relations

$$\int \mathsf{K}_{(\mathcal{A},\lambda)}(y)q(y_1)dy = (2\lambda^{\frac{1}{a}})^{-1} \int_{-\lambda^{\frac{1}{a}}}^{\lambda^{\frac{1}{a}}} q(y_1)dy_1,$$

and

$$\int \mathsf{K}_{(\mathcal{A},\lambda)}(y)Q(y_2, ..., y_d)dy = (2\lambda^{\frac{1}{b}})^{-(d-1)} \int_{-\lambda^{\frac{1}{b}}}^{\lambda^{\frac{1}{b}}} ... \int_{-\lambda^{\frac{1}{b}}}^{\lambda^{\frac{1}{b}}} Q(y_2, ..., y_d)dy_2...dy_d.$$

These properties are crucial to guarantee that the bias of kernel approximation is of the order $O(\lambda)$ (cf. Lemma 3 below). Note that the simple rectangular kernel (22) used for the local single index model can attain such a bias, but only at the price of too large L_2-norm (which characterizes the variance). We now give an example showing how a kernel with the required properties can be constructed in a particular case.

The two-step kernel Set

$$u_1 = \lambda^{\frac{1}{a}}, \quad u_2 = \lambda^{\frac{1}{b}}, \quad v_1 = \lambda^{\frac{b-a+1}{b^2}}, \quad v_2 = \frac{1}{2}\lambda^{\frac{1}{b}}, \tag{23}$$

$$\begin{aligned}
\Pi_{1,1} &= [0,u_1]\times[v_2,v_1]^{d-1}, \quad \mu_{1,1} = u_1(v_1-v_2)^{d-1};\\
\Pi_{2,2} &= [u_1,u_2]\times[0,v_2]^{d-1}, \quad \mu_{2,2} = (u_2-u_1)v_2^{d-1};\\
\Pi_{2,1} &= [u_1,u_2]\times[v_2,v_1]^{d-1}, \quad \mu_{2,1} = (u_2-u_1)(v_1-v_2)^{d-1}.
\end{aligned}$$

Next, we define, for $y \in \mathbb{R}^d_+$,

$$\Lambda(y) = \mu_{1,1}^{-1}\mathbb{I}_{\Pi_{1,1}}(y) - \mu_{2,1}^{-1}\mathbb{I}_{\Pi_{2,1}}(y) + \mu_{2,2}^{-1}\mathbb{I}_{\Pi_{2,2}}(y). \tag{24}$$

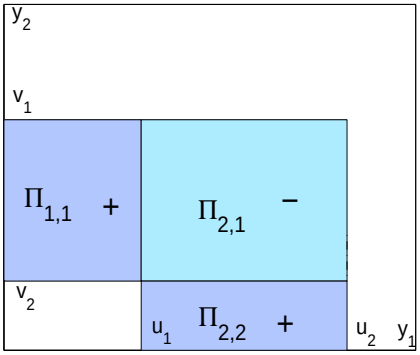

Figure 5. Pavement $\Pi_{i,j}$ for the two-step kernel, $d = 2$. The kernel vanishes in the blanc zones.

For $y = (y_1,\dots,y_d) \in \mathbb{R}^d$ we write $|y| = (|y_1|,\dots,|y_d|)$ and define the kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ for $y \in \mathbb{R}^d$ by the relation

$$\mathsf{K}_{(\mathcal{A},\lambda)}(y) = 2^{-d}\Lambda(|y|). \tag{25}$$

We will call this kernel the *two-step kernel* (cf. Figure 5). Its key property is as follows. First, for any integrable function $q(y_1)$ of the first coordinate y_1 we have

$$\int \mathsf{K}_{(\mathcal{A},\lambda)}(y)q(y_1)dy = \frac{1}{2u_1}\int_{-u_1}^{u_1} q(y_1)dy_1,$$

since the integral of q over $\Pi_{2,1}$ is exactly the same as that over $\Pi_{2,2}$. Further, for any integrable function $Q(y_2,...,y_d)$ of $y_2,...,y_d$,

$$\int \mathsf{K}_{(\mathcal{A},\lambda)}(y)Q(y_2,...,y_d)dy = (2v_2)^{-(d-1)}\int_{-v_2}^{v_2}\dots\int_{-v_2}^{v_2} Q(y_2,...,y_d)dy_2...dy_d,$$

since the integral of Q over $\Pi_{2,1}$ is exactly the same as that over $\Pi_{1,1}$. In words, the negative term $-\mu_{2,1}^{-1}\mathbb{I}_{\Pi_{2,1}}(y)$ in (24) allows us to compensate the excess of the bias introduced by the two other terms, so that the resulting bias remains of the order $O(\lambda)$ (cf. Lemma 3 below).

For the two-step kernel (25) we have

$$\int \mathsf{K}_{(\mathcal{A},\lambda)}(y)dy = 1, \quad \|\mathsf{K}_{(\mathcal{A},\lambda)}\|_1 = 3, \quad \|\mathsf{K}_{(\mathcal{A},\lambda)}\|_2^2 = \mu_{1,1}^{-1} + \mu_{2,2}^{-1} + \mu_{2,1}^{-1}.$$

We now define

$$\rho = \frac{(d-1)(a-1)}{b}$$

and consider the subset $\{\mathcal{A} = (a,b) : \rho \geq (b-a)/a\}$ of $\mathcal{P}_3$. It is easy to see that for $\rho \geq (b-a)/a$ we have

$$\|\mathsf{K}_{(\mathcal{A},\lambda)}\|_2^2 = O(\lambda^{-\frac{d}{b}}).$$

Since $a \leq b$ for $\mathcal{A} \in \mathcal{P}_3$, this result is better than part (ii) of Lemma 2 where $\mathsf{K}_{(\mathcal{A},\lambda)}$ is a rectangular kernel. But we need the condition $\rho \geq (b-a)/a$. It is clearly satisfied when $\rho \geq 1$ (recall that $a > 1, b \leq 2$). For smaller values of ρ we need to add extra "steps" in the construction, i.e., to introduce piecewise constant kernels with more and more pieces of the pavement, in order to get the bias compensation property as discussed above. For instance, if $\rho + \rho^2 \geq \frac{b-a}{a}$ (since $(b-a)/a < 1$, this is certainly the case when $\rho \geq \frac{\sqrt{5}-1}{2}$) we need a pavement of five sets $\Pi_{i,j}$ in order to obtain a piecewise constant kernel with the required statistical properties, and so on. We come to the following construction of the kernel.

Generic construction Define a piecewise constant kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ as follows. Fix an integer r that we will further call *number of steps* (of kernel construction). Let $(u_j)_{j=1,\dots,r}$ and $(v_j)_{j=1,\dots,r+1}$ be, respectively, a monotone increasing and a monotone decreasing sequence of positive numbers with $u_1 = \lambda^{\frac{1}{a}}$, $v_r = \lambda^{\frac{1}{b}}/2$ and $v_{r+1} = 0$. We set

$$\Pi_{1,1} = [0, u_1] \times [v_2, v_1]^{d-1}, \quad \mu_{1,1} = u_1(v_1 - v_2)^{d-1}.$$

For $i = 2, \dots, r$ and $j = i-1, i$ we define

$$\Pi_{i,j} = [u_{i-1}, u_i] \times [v_{j+1}, v_j]^{d-1}, \quad \mu_{i,j} = (u_i - u_{i-1})(v_j - v_{j+1})^{d-1}.$$

For $y \in \mathbb{R}^d_+$ consider

$$\begin{aligned} \Lambda_1(y) &= \frac{1}{\mu_{1,1}} \mathbb{I}_{\Pi_{1,1}}(y); \\ \Lambda_i(y) &= \frac{1}{\mu_{i,i}} \mathbb{I}_{\Pi_{i,i}}(y) - \frac{1}{\mu_{i,i-1}} \mathbb{I}_{\Pi_{i,i-1}}(y), \quad i = 2, ..., r. \end{aligned}$$

The kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ is defined for $y = (y_1, \dots, y_d) \in \mathbb{R}^d$ as follows:

$$\mathsf{K}_{(\mathcal{A},\lambda)}(y) = 2^{-d} \sum_{i=1}^{r} \Lambda_i(|y|) \tag{26}$$

where $|y| = (|y_1|, \dots, |y_d|)$. Clearly,

$$\int \mathsf{K}_{(\mathcal{A},\lambda)}(y) dy = 1, \quad \|\mathsf{K}_{(\mathcal{A},\lambda)}\|_1 = 2r - 1.$$

Construction of the kernel for $\mathcal{A} \in \mathcal{P}_3 = \{\mathcal{A} : 1 < a \leq b \leq 2,\ b < d(a-1)+1\}$ If $\rho \geq \frac{b-a}{a}$ we define $\mathsf{K}_{(\mathcal{A},\lambda)}$ as a two-step kernel, i.e., we set $r = 2$ and take (u_j) and (v_j) as in (23).

If $\rho < \frac{b-a}{a}$ we use another definition. We introduce the sequence $(\alpha_k)_{k \geq 0}$ as follows:

$$\alpha_0 = b^{-1}, \quad \alpha_{k+1} = \alpha_k \rho + b^{-1} = b^{-1} \sum_{i=0}^{k+1} \rho^i, \quad k = 1, 2, \dots \ . \tag{27}$$

The sequence (α_k) is monotone increasing and, since $b < d(a-1)+1$, we have

$$\lim_{k\to\infty} \alpha_k = \infty \quad \text{if} \quad \rho \geq 1, \quad \lim_{k\to\infty} \alpha_k = (b-(a-1)(d-1))^{-1} > \frac{1}{a} \quad \text{if} \quad \rho < 1. \tag{28}$$

Thus we can define an integer $r \geq 2$ such that

$$\alpha_{r-1} \geq \frac{1}{a} > \alpha_{r-2}\,. \tag{29}$$

Note that r depends only on $\mathcal{A} = (a,b)$ and d. Now we set

$$\begin{aligned} &u_1 = \lambda^{\frac{1}{a}}, \;\; u_i = \lambda^{\alpha_{r-i}}, \; i = 2,\dots,r; \\ &v_i = \lambda^{\frac{1}{b}} u_{i+1}^{-\frac{a-1}{b}}, \quad i = 1,\dots,r-1. \end{aligned} \tag{30}$$

Recall that $v_r = \frac{1}{2}\lambda^{\frac{1}{b}}$ and $v_{r+1} = 0$. If $\rho < \frac{b-a}{a}$ define the kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ by (26), with the sequences (u_j) and (v_j) as in (30).

Properties of the kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ Let $q : \mathbb{R} \to \mathbb{R}$ and $p : \mathbb{R}^d \to \mathbb{R}$, $B : \mathbb{R}^d \to \mathbb{R}$ be functions such that p is continuously differentiable and, for given $\mathcal{A} = (a,b) \in \mathcal{P}_2 \cup \mathcal{P}_3$ and $\lambda > 0$,

$$\Big| q(0) - \frac{1}{2\lambda^{1/a}} \int_{-\lambda^{1/a}}^{\lambda^{1/a}} q(z)dz \Big| \leq c_3\lambda, \tag{31}$$

$$\big| p(z') - p(z) - [\nabla p(z)]^T (z'-z) \big| \leq L\|z'-z\|^b, \quad \forall\, z, z' \in \mathbb{R}^d, \tag{32}$$

$$\sup_{x\in\mathbb{R}^d} |B(x)| \leq c_4 \tag{33}$$

where c_3, c_4 and L are positive constants. Let $\mathfrak{B}(\mathcal{A},\lambda)$ denote the set of triplets (q,p,B) satisfying (31) – (33). Define

$$Q(y) = q(y_1) + p(y) + B(y)|y_1|^{a-1}\|y\|^b, \quad \forall\; y \in \mathbb{R}^d.$$

Lemma 3. *Let $\mathcal{A} = (a,b) \in \mathcal{P}_3$. Let the kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ be defined by (26), with the sequences (u_i) and (v_i) as in (30) if $\rho < \frac{b-a}{a}$, and with $r = 2$, (u_i) and (v_i) as in (23) if $\rho \geq \frac{b-a}{a}$. Then, for any $\lambda > 0$ small enough,*

$$\sup_{(q,p,B)\in\mathfrak{B}(\mathcal{A},\lambda)} \big| \big[\mathsf{K}_{(\mathcal{A},\lambda)} * Q\big](0) - Q(0) \big| \leq c\lambda, \tag{34}$$

$$\int \big|\mathsf{K}_{(\mathcal{A},\lambda)}(y)\big| \, \big\|y\|^m du \leq c'\lambda^{\frac{m}{ab}}, \;\; \forall m \in \mathbb{R}, \tag{35}$$

where the constant c depends only on c_3, c_4, L, d and $\mathcal{A}$, and c' depends only on m, d and $\mathcal{A}$. Furthermore,

$$\|\mathsf{K}_{(\mathcal{A},\lambda)}\|_1 \leq c'' \quad \textit{and} \quad \|\mathsf{K}_{(\mathcal{A},\lambda)}\|_2 \leq c^{(3)}\lambda^{-\frac{d}{2b}} \tag{36}$$

where the constants c'' and $c^{(3)}$ only depend on $\mathcal{A}$ and d.

Note that for $\rho \geq \frac{b-a}{a}$ the kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ in this lemma is just the two-step kernel. The corresponding pavement $\{\Pi_{i,j}\}$ only contains three sets (cf. Figure 5).

The kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ depends on $\mathcal{A} = (a,b)$ in such a way that the constants in the bounds (34) – (36) diverge when $\mathcal{A}$ approaches the boundary $d(a-1)+1 = b$ of the zone $\mathcal{P}_3$. So, Lemma 3 cannot be extended to $\mathcal{A} \in \mathcal{P}_2$.

Construction of the kernel for $\mathcal{A} \in \mathcal{P}_2$ We consider now another choice of the sequences (u_i) and (v_i) which provides the kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ with the properties similar to those of Lemma 3 but satisfied for all $\mathcal{A} \in \mathcal{P}_2 \cup \mathcal{P}_3$ and, what is more, uniformly over this set. The price to pay for the uniformity is an extra $\log\log(1/\lambda)$ factor in the bound for the L_1-norm of $\mathsf{K}_{(\mathcal{A},\lambda)}$.

If $(b-a)/a \le (1+\rho)\rho$ we define the kernel as in Lemma 3. If $(b-a)/a > (1+\rho)\rho$ we use another definition of sequences (u_i) and (v_i). For any $0 < \lambda < 1$ we define

$$V(\lambda) = \ln\left\{\frac{(a-1)(b-a)}{ab^2}\ln(1/\lambda)\right\}. \tag{37}$$

If $V(\lambda) \le 0$ we define $\mathsf{K}_{(\mathcal{A},\lambda)}$ as a two-step kernel, i.e., we set $r = 2$ and take (u_j) and (v_j) as in (23). If $V(\lambda) > 0$ we define $r = r(\lambda) > 1$ by

$$r = \min\left\{s \in \mathbb{N} :\ s > 1,\ \frac{V(\lambda)}{s-1} < \frac{1}{2}\ln\left(\frac{\sqrt{5}+1}{2}\right)\right\}.$$

Next, set $\alpha = \frac{V(\lambda)}{r-1}$, $\nu = \left(\frac{\sqrt{5}+1}{2}\right)^{1/2}$ and define the sequences (u_i) and (v_i) as follows

$$\begin{aligned} &u_i = \lambda^{\frac{1}{a}}\exp\left\{\tfrac{b}{a-1}\exp(\alpha(i-1))\right\},\ \ i = 1, ..., r,\\ &v_i = \lambda^{\frac{1}{ab}}\exp\left\{-\nu\exp(\alpha i)\right\},\ i = 1, \dots, r-1, \quad v_r = \tfrac{1}{2}\lambda^{\frac{1}{b}}. \end{aligned} \tag{38}$$

Note that $u_r = \lambda^{\frac{1}{b}}$.

Lemma 4. *Let $\mathcal{A} = (a,b) \in \mathcal{P}_2 \cup \mathcal{P}_3$. Let the kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ be defined by (26), with the sequences (u_i), (v_i) as in Lemma 3 if $(1+\rho)\rho \ge \frac{b-a}{a}$, with the sequences (u_i), (v_i) as in (38) if $(1+\rho)\rho < \frac{b-a}{a}$ and $V(\lambda) > 0$, and with $r = 2$, (u_i) and (v_i) as in (23) if $(1+\rho)\rho < \frac{b-a}{a}$ and $V(\lambda) \le 0$. Then, for any $\lambda > 0$ small enough,*

$$\sup_{(q,p,B)\in\mathfrak{B}(\mathcal{A},\lambda)} \left|\left[\mathsf{K}_{(\mathcal{A},\lambda)} * Q\right](0) - Q(0)\right| \le c_5\lambda, \tag{39}$$

$$\int \left|\mathsf{K}_{(\mathcal{A},\lambda)}(y)\right| \|y\|^m\, du \le c_6\lambda^{\frac{m}{ab}},\ \forall m \in \mathbb{R}, \tag{40}$$

where the constant c_5 depends only on c_3, c_4, L and d, and $c_6 > 0$ depends only on m and d (both constants are explicit in the proof of the lemma). Furthermore,

$$\|\mathsf{K}_{(\mathcal{A},\lambda)}\|_1 \le c_7\ln\ln\lambda^{-1} \quad \textit{and} \quad \|\mathsf{K}_{(\mathcal{A},\lambda)}\|_2 \le c_8\lambda^{-\frac{b+d-1}{2ab}} \tag{41}$$

where the constants c_7 and c_8 only depend on d.

Some remarks are in order here.

1. The number of steps r in the construction of the kernel is typically small. In particular, $r = 2$ if $\rho \ge \frac{b-a}{a}$, and $r = 3$ if $(1+\rho)\rho \ge \frac{b-a}{a} > \rho$ (cf. (29)). Moreover, for $1 < a \le b \le 2$ we have
$$\frac{(a-1)(b-1)}{ab^2} \le \frac{(b-1)^2}{b^3} \le \frac{1}{8}.$$
Hence, $V(\lambda) \le \ln\left(\frac{\sqrt{5}+1}{2}\right)$ for all $\lambda > 3\cdot 10^{-6}$ which means that, for $(1+\rho)\rho < \frac{b-a}{a}$, no more than 3 steps of the construction are needed if $\lambda > 3\cdot 10^{-6}$. In other words, unless we are not "extremely far" in the asymptotics, the number of steps r does not exceed 3 and thus the L_1-norm of the resulting kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ is bounded by 5.

2. In the asymptotics when $\lambda \to 0$ the number of steps $r = r(\lambda)$ in the construction and thus the L_1-norm of the kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ is at most $O(\ln\ln\lambda^{-1})$. As discussed in the previous remark, this behavior starts "extremely far" in the asymptotics, so it has essentially a theoretical interest. In the theory, it results in an extra $\ln\ln\varepsilon^{-1}$ factor in the upper bound for the adaptive estimation procedure, as compared to the lower bound in (7). It can be shown that for $\mathcal{A} \in \mathcal{P}_2$ a kernel with the required approximation properties cannot have the L_1-norm growing slower than $\ln\ln\lambda^{-1}$, as $\lambda \to 0$. On the other hand, as we have seen in Lemma 3, for $\mathcal{A} \in \mathcal{P}_3$ solely, there is a choice of sequences (u_j) and (v_j) such that the L_1-norm of the kernel is bounded by a constant independent of λ. This constant, however, depends on $\mathcal{A} = (a, b)$ and explodes as $\mathcal{A}$ approaches the boundary of $\mathcal{P}_3$.

7.3 Basic approximation results

We can now describe the approximation properties of the kernel $K_{\mathcal{J}}$ which serve as a main tool in the proof of the properties of the estimator $g^*_{\mathcal{A},\varepsilon}(x)$.

Let $x \in [-1,1]^d$ and $\mathcal{A}_0 = (\gamma, \beta) \in (0,2]^2$ be fixed and let $g = f \circ G \in \mathbb{H}(\mathcal{A}_0, \mathcal{L})$. We define

$$\vartheta_0^x \triangleq \begin{cases} (1,0,\dots,0) & \text{if} \quad \beta > 1 \text{ and } \nabla G(x) = 0, \text{ or } \beta \le 1, \\ \nabla G(x)/\|\nabla G(x)\| & \text{if} \quad \beta > 1,\ \nabla G(x) \neq 0. \end{cases} \tag{42}$$

The following statement is an immediate consequence of Lemmas 1 – 4.

Corollary 1. *For all $\mathcal{A}_0 = (\gamma, \beta) \in (0,2]^2$, and all $\lambda > 0$ we have*

$$\sup_{x\in[-1,1]^d} \sup_{g\in\mathbb{H}(\mathcal{A}_0,\mathcal{L})} \left|\left[K_{\mathcal{J}_0^x} * g\right](x) - g(x)\right| \le c_{10}\lambda$$

where $\mathcal{J}_0^x = \left(\mathcal{A}_0, \vartheta_0^x, \lambda\right)$ and c_{10} is a constant depending only on $\mathcal{A}_0$, $\mathcal{L}$ and d.

In other words, the collection $\{K_{\mathcal{J}},\ \mathcal{J} \in \mathfrak{J}\}$ of the kernels contains an element $K_{\mathcal{J}_0^x}$ such that the quality of approximation of $g(x)$ by the "ideal" smoother $\left[K_{\mathcal{J}_0^x} * g\right](x)$ is of the order $O(\lambda)$. Here we use the term "ideal" because $\mathcal{J}_0^x = (\mathcal{A}_0, \vartheta_0^x, \lambda)$ depends on the gradient $\nabla G(x)$, and thus on the unknown function g.

In what follows we also need another property of kernels $K_{\mathcal{J}}$.

Proposition 1. *For all $\mathcal{A}_0 = (\gamma, \beta) \in (0,2]^2$, $x \in [-1,1]^d$, $\lambda_0 > 0$ and all $\mathcal{J} = \left(\mathcal{A}, \vartheta, \lambda\right) \in \mathfrak{J}$ such that $\lambda^{\frac{1}{a\flat}} \le 2\lambda_0^{\frac{1}{\gamma\beta}} \wedge 1$ we have*

$$\sup_{\mathcal{A}\in(0,2]^2} \sup_{g\in\mathbb{H}(\mathcal{A}_0,\mathcal{L})} \left|\left[\Delta_{\mathcal{J}} K_{\mathcal{J}*\mathcal{J}_0^x} * g\right](x)\right| \le c_{11}\left\{\left(\|K_{\mathcal{J}}\|_1 + \|K_{\mathcal{J}_0^x}\|_1\right)\lambda_0 + \|K_{\mathcal{J}}\|_1\|K_{\mathcal{J}_0^x}\|_1\,\varepsilon^2\right\}, \tag{43}$$

where $\mathcal{J}_0^x = \left(\mathcal{A}_0, \vartheta^x, \lambda_0\right)$, ϑ^x is any element of the unit sphere $\mathbb{S}$ such that $\|\vartheta^x - \vartheta_0^x\| \le \varepsilon^2$ and c_{11} is a constant depending only on $\mathcal{A}_0$, $\mathcal{L}$ and d. Furthermore, for any $\mathcal{J}, \mathcal{J}' \in \mathfrak{J}$ we have

$$\|\Delta_{\mathcal{J}'} K_{\mathcal{J}'*\mathcal{J}}\|_2 \le \Big(\|K_{\mathcal{J}'}\|_1 + \|K_{\mathcal{J}}\|_1\Big)\|K_{\mathcal{J}'}\|_2. \tag{44}$$

8 Proof of Theorem 1

For any $\beta > 0, \gamma > 0$ and any $0 < \varepsilon < 1$ define the integers

$$q_1 = \lceil \Big(\varepsilon \sqrt{\ln(1/\varepsilon)} \Big)^{-\frac{2}{2\gamma\beta+\beta+(d-1)}} \rceil .$$

Consider the regular grid Γ_{q_1} on $[0,1]^{d-1}$ defined by

$$\Gamma_{q_1} \triangleq \left\{ \left(\frac{2k_1+1}{2q_1}, \ldots, \frac{2k_{d-1}+1}{2q_1} \right) : k_i \in \{0, \ldots, q_1 - 1\}, i = 1, \ldots, d-1 \right\} .$$

Denote by $x_1, \ldots, x_m$, where $m = \mathrm{card}(\Gamma_{q_1}) = q_1^{d-1}$, the elements of Γ_{q_1} numbered in an arbitrary order.

Let $u : \mathbb{R} \to \mathbb{R}_+$ be an infinitely differentiable function such that $u(0) = 1, u(t) = u(-t)$ for all $t \in \mathbb{R}$, $\mathrm{supp}\, u = [-1/2, 1/2]$ and $u(t)$ is strictly monotone decreasing on $[0, 1/2]$. Set $f_0(t) = u(t), \forall\ t \in \mathbb{R}$, and

$$\varphi_0(t_2, \ldots, t_d) = \frac{1}{2} \prod_{j=2}^{d} u(t_j), \quad \forall\ t \in \mathbb{R}.$$

Define the following infinitely differentiable functions of $t = (t_1, \ldots, t_d) \in [-1,1]^d$:

$$\begin{aligned} g_0(t) &= L_0 h^\gamma f_0\Big(\frac{t_1}{h}\Big), \\ g_k(t) &= L_0 h^\gamma f_0\Big(\frac{t_1}{h} + \frac{L_0 h_1^\beta}{h} \varphi_0\Big(\frac{t_2 - x_{k,2}}{h_1}, \ldots, \frac{t_d - x_{k,d}}{h_1}\Big)\Big), \quad k = 1, \ldots, m, \end{aligned}$$

where $h = h_1^\beta,\ h_1 = 1/q_1,\ 0 < L_0 < 1$ is a constant to be chosen small enough, and $x_{k,j}$ stands for the jth component of x_k. We note that, in view of the above definitions, the sets where the functions g_l and g_k differ from g_0 are disjoint for $l \neq k,\ k \neq 0,\ l \neq 0$.

It is easy to see that if L_0 is small enough, $g_k \in \mathbb{H}(\mathcal{A}, \mathcal{L})$, $k = 0, \ldots, m$. In what follows, we assume that L_0 is chosen in this way. To prove Theorem 1, we follow the scheme of proving lower bounds based on reduction to the problem of distinguishing between $m+1$ hypotheses (cf., e.g., (25)). We choose the hypotheses to be determined by $g_0, \ldots, g_m$ and we apply Theorem 2.5 of (25), where we consider the sup-norm distance $d(g_l, g_k) = \|g_l - g_k\|_\infty = \sup_{t \in [-1,1]^d} |g_l(t) - g_k(t)|$, $l, k = 1, \ldots, m$. Since the functions g_l and g_k differ from g_0 on disjoint sets, for any $l \neq k,\ l, k = 1, \ldots, m$, we have

$$\begin{aligned} d(g_l, g_k) = d(g_0, g_k) &\geq L_0 h^\gamma |f_0(0) - f_0(L_0 h_1^\beta \varphi_0(0)/h)| \\ &= L_0 h^\gamma |f_0(0) - f_0(L_0(1 + o_\varepsilon(1))/2)|, \end{aligned}$$

where $o_\varepsilon(1) \to 0$, as $\varepsilon \to 0$. Since $L_0 > 0$ and f_0 is strictly decreasing on $[0, \infty)$, there exists a constant $L^* > 0$ such that, for ε small enough,

$$d(g_l, g_k) \geq L^* h^\gamma \asymp \Big(\varepsilon \sqrt{\ln(1/\varepsilon)} \Big)^{\frac{2\gamma}{2\gamma+1+(d-1)/\beta}}, \quad l \neq k,\ l, k = 0, \ldots, m. \tag{45}$$

Thus, assumption (i) of Theorem 2.5 in (25) is satisfied with $s = L^* h^\gamma / 2$. It remains to check assumption (ii) of that theorem. The probability measures $\mathbb{P}_{g_k}$ are gaussian, and the

Kullback-Leibler divergence between $\mathbb{P}_{g_k}$ and $\mathbb{P}_{g_0}$ has the form

$$\begin{aligned}\mathbf{K}(\mathbb{P}_{g_k},\mathbb{P}_{g_0}) &= \varepsilon^{-2}\int_{\mathcal{D}}(g_0(t)-g_k(t))^2dt\\ &= \varepsilon^{-2}L_0^2h^{2\gamma}\int_{\mathcal{D}}\Big|f_0\Big(\frac{t_1}{h}\Big)-f_0\Big(\frac{t_1}{h}+w(t_2,\dots,t_d)\Big)\Big|^2dt\end{aligned}$$

where we write for brevity

$$w(t_2,\dots,t_d)\triangleq\frac{L_0h_1^\beta}{h}\varphi_0\Big(\frac{t_2-x_{k,2}}{h_1},\dots,\frac{t_d-x_{k,d}}{h_1}\Big).$$

Since, for any $w\in\mathbb{R}$,

$$\Big|f_0\Big(\frac{t_1}{h}\Big)-f_0\Big(\frac{t_1}{h}+w\Big)\Big|^2=w^2\Big|\int_0^1 f_0'\Big(\frac{t_1}{h}+uw\Big)du\Big|^2\le w^2\int_0^1\Big|f_0'\Big(\frac{t_1}{h}+uw\Big)\Big|^2du$$

we find

$$\begin{aligned}\mathbf{K}(\mathbb{P}_{g_k},\mathbb{P}_{g_0}) &\le \varepsilon^{-2}L_0^2h^{2\gamma}\int w^2(t_2,\dots,t_d)dt_2\dots dt_d\times\\ &\qquad\int_0^1\left[\int\Big|f_0'\Big(\frac{t_1}{h}+uw(t_2,\dots,t_d)\Big)\Big|^2dt_1\right]du\\ &= L_0^4\varepsilon^{-2}h^{2\gamma+1}h_1^{d-1}\int_{\mathbb{R}^{d-1}}\varphi_0^2(v)dv\int_{\mathbb{R}}|f_0'(v_1)|^2dv_1\\ &\le c_*L_0^4\ln(1/\varepsilon)\end{aligned}$$

where $c_*>0$ is an absolute constant. Next, $m=q_1^{d-1}$, so that $\ln m\asymp\ln(1/\varepsilon)$. This and the previous inequality imply that if L_0 is chosen small enough, we have

$$\mathbf{K}(\mathbb{P}_{g_k},\mathbb{P}_{g_0})\le(1/16)\ln m. \tag{46}$$

Using (45), (46) and applying Theorem 2.5 in (25) we get the lower bound

$$\liminf_{\varepsilon\to0}\inf_{\tilde g_\varepsilon}\sup_{g\in\mathbb{H}(\mathcal{A},\mathcal{L})}\mathbb{E}_g\left[\left(\big(\varepsilon\sqrt{\ln(1/\varepsilon)}\big)^{-\frac{2\gamma}{2\gamma+1+(d-1)/\beta}}\|\tilde g_\varepsilon-g\|_\infty\right)^p\right]>0, \tag{47}$$

which is valid for all $\beta>0,\gamma>0$ and all $p>0$.

We now show that for the trivial cases discussed in Section 2 we can obtain better lower bounds. Consider first the case where $0<\beta,\gamma\le1$. Then we use the same technique as above, but we set now $q_1=\lceil\Big(\varepsilon\sqrt{\ln(1/\varepsilon)}\Big)^{-\frac{2}{2\gamma\beta+d}}\rceil$. We then introduce a regular grid $\Gamma^*_{q_1}$ on $[0,1]^d$ defined by

$$\Gamma^*_{q_1}\triangleq\left\{\left(\frac{2k_1+1}{2q_1},\dots,\frac{2k_d+1}{2q_1}\right):k_i\in\{0,\dots,q_1-1\},i=1,\dots,d\right\}$$

and denote by $x_1,\dots,x_m$, where $m=\mathrm{card}(\Gamma^*_{q_1})=q_1^d$, the elements of $\Gamma^*_{q_1}$ numbered in an arbitrary order. We set

$$\varphi_0(t)\triangleq\prod_{j=1}^d u(t_j),\quad\forall\ t\in\mathbb{R},$$

and we choose the functions g_k in a different way:

$$\begin{aligned} g_0(t) &= |t|^\gamma, \\ g_k(t) &= \Big|t + L_0 h^\beta \varphi_0\Big(\frac{t - x_k}{h}\Big)\Big|^\gamma, \quad k = 1, \dots, m, \end{aligned}$$

where $h = 1/q_1$. With this choice, clearly,

$$d(g_l, g_k) \geq L_0^\gamma h^{\gamma\beta} \varphi_0^\gamma(0) \asymp \Big(\varepsilon\sqrt{\ln(1/\varepsilon)}\Big)^{\frac{2\gamma\beta}{2\gamma\beta+d}}, \quad l \neq k,\ l, k = 0, \dots, m. \tag{48}$$

Next,

$$\begin{aligned} \mathbf{K}(\mathbb{P}_{g_k}, \mathbb{P}_{g_0}) &= \varepsilon^{-2} \int_{\mathcal{D}} (g_0(t) - g_k(t))^2 dt \\ &\leq L_0^{2\gamma} \varepsilon^{-2} h^{2\gamma\beta+d} \int_{\mathbb{R}^d} \varphi_0^{2\gamma}(v) dv \\ &= O\big(\ln(1/\varepsilon)\big), \quad \text{as } \varepsilon \to 0. \end{aligned} \tag{49}$$

Using (48), (49) and Theorem 2.5 in (25), the proof is completed as in the previous case, so that we get the lower bound

$$\liminf_{\varepsilon\to 0} \inf_{\tilde{g}_\varepsilon} \sup_{g \in \mathbb{H}(\mathcal{A},\mathcal{L})} \mathbb{E}_g \left[\left(\big(\varepsilon\sqrt{\ln(1/\varepsilon)}\big)^{-\frac{2\gamma\beta}{2\gamma\beta+d}} \|\tilde{g}_\varepsilon - g\|_\infty \right)^p \right] > 0, \tag{50}$$

which is valid for all $0 < \beta, \gamma \leq 1$ and all $p > 0$.

Finally, the second trivial case where (47) can be improved corresponds to $\gamma \geq \beta \vee 1$. As observed in Section 2, in this case we have the inclusion $\mathbb{H}_d(\beta, L_4) \subset \mathbb{H}(\mathcal{A}, \mathcal{L})$ with some constant $L_4 > 0$, and we can use the standard lower bound for $\mathbb{H}_d(\beta, L_4)$ (cf. (20; 5; 3)):

$$\liminf_{\varepsilon\to 0} \inf_{\tilde{g}_\varepsilon} \sup_{g \in \mathbb{H}(\mathcal{A},\mathcal{L})} \mathbb{E}_g \left[\left(\big(\varepsilon\sqrt{\ln(1/\varepsilon)}\big)^{-\frac{2\beta}{2\beta+d}} \|\tilde{g}_\varepsilon - g\|_\infty \right)^p \right] > 0. \tag{51}$$

Combining the bounds (47), (50) and (51) we obtain the result of Theorem 1.

9 Proof of Theorem 2

We need the following technical result.

Lemma 5. *Let $\zeta = (\zeta_1, \dots, \zeta_\mathcal{M})$ be a gaussian random vector defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and such that $\mathbb{E}\zeta_m = 0$, $\mathbb{E}\zeta_m^2 = \sigma_m^2$, $m = 1, \dots, \mathcal{M}$. Let $\mathfrak{m}$ be a random variable with the values in $(1, \dots, \mathcal{M})$ defined on the same probability space. Then for all $A > 1$ and all $s > 0$ we have*

$$\mathbb{E}\left(|\zeta_\mathfrak{m}|^s\right) \leq \Big(\sqrt{2A\ln(\mathcal{M})}\Big)^s \Big\{ \mathbb{E}\big(\sigma_\mathfrak{m}^s\big) + c_{12}(A, s) \mathcal{M}^{1-A} \max_{m=1,\dots,\mathcal{M}} \sigma_m^s \Big\}$$

where $c_{12}(A, s) > 0$ is a constant depending only on A and s.

Proof is standard (see, e.g., (11)).

To prove Theorem 2 we proceed in steps.

1°.*Reduction to the discrete norm.* Fix $\mathcal{A} = (\gamma, \beta) \in (0, 2]^2$, and suppose that $g \in \mathbb{H}(\mathcal{A}, \mathcal{L})$. Let, for brevity, $\bar{g}_\varepsilon^* = g_{\mathcal{A},\varepsilon}^*$. In view of the construction of the global estimator (cf. (21)) we get, for all $g \in \mathbb{H}(\mathcal{A}, \mathcal{L})$,

$$\begin{aligned} \|\bar{g}_\varepsilon^* - g\|_\infty &\le \sup_{z\in\mathbb{Z}^d} \max_{x\in\Pi_\varepsilon(z)\cap[-1,1]^d} \left|\bar{g}_\varepsilon^*(x) - g(x)\right| \\ &\le |\bar{g}_\varepsilon^* - g|_\infty + C\varepsilon^{2\gamma(\beta\wedge 1)} \end{aligned} \tag{52}$$

where

$$|\bar{g}_\varepsilon^* - g|_\infty \triangleq \max_{z\in\mathcal{Z}_\varepsilon} \left|\bar{g}_\varepsilon^*(z) - g(z)\right| \quad \text{with} \quad \mathcal{Z}_\varepsilon = \left(\varepsilon^2\mathbb{Z}\right)^d \cap [-1, 1]^d.$$

Here and in what follows we will use the same notation C for possibly different positive constants depending only on $\mathcal{A}, \mathcal{L}$ and d. Since $\varepsilon^{2\gamma(\beta\wedge 1)} = o(\phi_\varepsilon(\gamma, \beta))$, $\varepsilon \to 0$, for all $(\gamma, \beta) \in \mathbb{R}_+^2$, it is sufficient to prove Theorem 2 with the loss given by the maximum norm $|\cdot|_\infty$ on the finite set $\mathcal{Z}_\varepsilon$. *Thus, w.l.o.g. in what follows we will replace* $\|\cdot\|_\infty$ *by* $|\cdot|_\infty$.

2°.*Control of large deviations.* To any $z \in \mathcal{Z}_\varepsilon$ we assign a vector $\theta^z \in \mathbb{S}_\varepsilon$ such that $\|\theta^z - \theta_0^z\| \le \varepsilon^2$ where θ_0^z is defined in (42). Next, we set $\mathcal{J}_0^z \triangleq (\mathcal{A}, \theta^z, \lambda_\varepsilon(\mathcal{A}))$. Introduce the random event

$$\mathcal{F} = \left\{\exists z \in \mathcal{Z}_\varepsilon : \ \mathcal{J}_0^z \notin \hat{\mathfrak{T}}_z\right\},$$

where $\hat{\mathfrak{T}}_z$ is the set of acceptable triplets $\mathcal{J}$ defined in Section 6. We now show that for all $\varepsilon > 0$ small enough

$$\sup_{g\in\mathbb{H}(\mathcal{A},\mathcal{L})} \mathbb{P}_g(\mathcal{F}) \le c_{12}\varepsilon^{2p} \tag{53}$$

where the constant c_{12} depends only on d. Indeed, in view of the definition of the random set $\hat{\mathfrak{T}}_z$,

$$\mathcal{F} \subseteq \bigcup_{z\in\mathcal{Z}_\varepsilon} \bigcup_{\mathcal{J}'\in\mathfrak{J}_{\mathrm{grid}}} \left\{\left|\Delta_{\mathcal{J}'}\hat{g}_{\mathcal{J}'*\mathcal{J}_0^z}(z)\right| > \mathbf{TH}_\varepsilon\left(\mathcal{J}', \mathcal{J}_0^z\right)\right\}$$

and therefore

$$\mathbb{P}_g(\mathcal{F}) \le \sum_{z\in\mathcal{Z}_\varepsilon} \sum_{\mathcal{J}'\in\mathfrak{J}_{\mathrm{grid}}} \mathbb{P}_g\left\{\left|\Delta_{\mathcal{J}'}\hat{g}_{\mathcal{J}'*\mathcal{J}_0^z}(z)\right| > \mathbf{TH}_\varepsilon\left(\mathcal{J}', \mathcal{J}_0^z\right)\right\}. \tag{54}$$

Note that

$$\mathbb{E}_g\Delta_{\mathcal{J}'}\hat{g}_{\mathcal{J}'*\mathcal{J}_0^z}(z) = \left[\Delta_{\mathcal{J}'}K_{\mathcal{J}'*\mathcal{J}_0^z} * g\right](z).$$

Applying Proposition 1 with $\mathcal{A}_0 = \mathcal{A}$, $\mathcal{J}_0^z = (\mathcal{A}, \theta^z, \lambda_\varepsilon(\mathcal{A}))$ and $\lambda = \lambda_0 = \lambda_\varepsilon(\mathcal{A})$ we obtain,

$$\begin{aligned} \sup_{g\in\mathbb{H}(\mathcal{A},\mathcal{L})} \left|\mathbb{E}_g\Delta_{\mathcal{J}'}\hat{g}_{\mathcal{J}'*\mathcal{J}_0^z}(z)\right| \le c_{11}\Big\{&\lambda_\varepsilon(\mathcal{A})\Big(\|K_{\mathcal{J}'}\|_1 + \|K_{\mathcal{J}_0^z}\|_1\Big) \\ &+ \|K_{\mathcal{J}'}\|_1\|K_{\mathcal{J}_0^z}\|_1\,\varepsilon^2\Big\} \end{aligned} \tag{55}$$

Now, due to the construction of the kernel $\mathsf{K}_{(\mathcal{A},\lambda)}$ and the fact that $\|K_{\mathcal{J}}\|_1 = \|\mathsf{K}_{(\mathcal{A},\lambda_\varepsilon(\mathcal{A}))}\|_1$ for all $\mathcal{J} \in \mathfrak{J}_{\mathrm{grid}}$, there exists a constant c_{13} depending only on $\mathcal{A}$ and d such that $K_{\mathcal{A}}^* \triangleq \max_{\mathcal{J}\in\mathfrak{J}_{\mathrm{grid}}} \|K_{\mathcal{J}}\|_1$ satisfies

$$\begin{aligned} K_{\mathcal{A}}^* &\le c_{13} \text{ if } \mathcal{A} \in (0, 2]^2 \setminus \mathcal{P}_2, \\ K_{\mathcal{A}}^* &\le c_{13}\ln\ln(1/\varepsilon) \text{ if } \mathcal{A} \in \mathcal{P}_2. \end{aligned}$$

Since also $\|K_{\mathcal{J}}\|_1 \geq 1$ and $\lambda_\varepsilon(\mathcal{A})/(\varepsilon \ln\ln(1/\varepsilon)) \to \infty$, as $\varepsilon \to 0$, we have, for $\varepsilon > 0$ small enough,

$$\begin{aligned}\sup_{g\in\mathbb{H}(\mathcal{A},\mathcal{L})} \left|\mathbb{E}_g \Delta_{\mathcal{J}'} \hat{g}_{\mathcal{J}'*\mathcal{J}_0^z}(z)\right| &\leq 2c_{11}\lambda_\varepsilon(\mathcal{A})\Big(\|K_{\mathcal{J}'}\|_1 + \|K_{\mathcal{J}_0^z}\|_1\Big) \\ &= 2\varepsilon\sqrt{\ln(1/\varepsilon)}\|\mathsf{K}_{(\mathcal{A},\lambda_\varepsilon(\mathcal{A}))}\|_2 \ \Big(\|K_{\mathcal{J}'}\|_1 + \|K_{\mathcal{J}_0^z}\|_1\Big)\end{aligned} \tag{56}$$

where we used that $\lambda_\varepsilon(\mathcal{A})$ is a solution of (19). Note also that in $\mathbb{P}_g$-probability

$$\Delta_{\mathcal{J}'}\hat{g}_{\mathcal{J}'*\mathcal{J}_0^z}(z) - \mathbb{E}_g \Delta_{\mathcal{J}'}\hat{g}_{\mathcal{J}'*\mathcal{J}_0^z}(z) \ \sim \ \mathcal{N}\Big(0, \varepsilon^2 \|\Delta_{\mathcal{J}'} K_{\mathcal{J}'*\mathcal{J}_0^z}\|_2^2\Big). \tag{57}$$

Using (44), (54) – (57) and the definition of the threshold $\mathbf{TH}_\varepsilon(\cdot,\cdot)$ we obtain that, for $\varepsilon > 0$ small enough,

$$\mathbb{P}_g(\mathcal{F}) \leq \mathrm{card}(\mathcal{Z}_\varepsilon)\mathrm{card}(\mathbb{S}_\varepsilon)\mathbb{P}\Big\{|\xi| > \sqrt{(4p+8d)\ln(1/\varepsilon)}\Big\} \leq \mathrm{card}(\mathcal{Z}_\varepsilon)\mathrm{card}(\mathbb{S}_\varepsilon)\varepsilon^{2p+4d}$$

where $\xi \sim \mathcal{N}(0,1)$. This proves (53) since $\mathrm{card}(\mathcal{Z}_\varepsilon) \leq (2\varepsilon^{-2}+1)^d$ and $\mathrm{card}(\mathbb{S}_\varepsilon) \leq (\sqrt{d}\varepsilon^{-2})^d$.

3°. *Two intermediate bounds on the risks.* Using that $|\bar{g}_\varepsilon^*| \leq \ln\ln(1/\varepsilon)$ and $g \in \mathbb{H}(\mathcal{A},\mathcal{L})$ is uniformly bounded we deduce from (53) that, for all $\mathcal{A} = (\gamma,\beta) \in (0,2]^2$,

$$\limsup_{\varepsilon\to 0} \sup_{g\in\mathbb{H}(\mathcal{A},\mathcal{L})} \mathbb{E}_g\Big(\phi_\varepsilon^{-p}(\gamma,\beta)|\bar{g}_\varepsilon^* - g|_\infty^p \mathbb{I}\{\mathcal{F}\}\Big) = 0. \tag{58}$$

We now control the bias of $\hat{g}_{\mathcal{J}_0^z}$ via Corollary 1, its stochastic error via the bounds on $\|\mathsf{K}_{(\mathcal{A},\lambda_\varepsilon(\mathcal{A}))}\|_2$ in Lemmas 2 – 4 and apply (19) to get that, for all $\mathcal{A} = (\gamma,\beta) \in (0,2]^2$,

$$\limsup_{\varepsilon\to 0} \sup_{g\in\mathbb{H}(\mathcal{A},\mathcal{L})} \mathbb{E}_g\Big(\phi_\varepsilon^{-p}(\gamma,\beta)|\hat{g}_{\mathcal{J}_0^z} - g|_\infty^p\Big) < \infty. \tag{59}$$

4°. *Final argument.* Note that on the event $\mathcal{F}^c$ the set $\hat{\mathfrak{T}}_z$ of acceptable triplets $\mathcal{J}$ is non-empty for every $z \in \mathcal{Z}_\varepsilon$, so that $\hat{\mathcal{J}}_z$ exists. Thus, on $\mathcal{F}^c$ we can write, for all $z \in \mathcal{Z}_\varepsilon$,

$$\Big|\hat{g}_{\hat{\mathcal{J}}_z}(z) - g(z)\Big| \leq \Big|\Delta_{\hat{\mathcal{J}}_z}\hat{g}_{\hat{\mathcal{J}}_z*\mathcal{J}_0^z}(z)\Big| + \Big|\Delta_{\mathcal{J}_0^z}\hat{g}_{\mathcal{J}_0^z*\hat{\mathcal{J}}_z}(z)\Big| + \Big|\hat{g}_{\mathcal{J}_0^z}(z) - g(z)\Big|. \tag{60}$$

Further, on $\mathcal{F}^c$ the triplet $\mathcal{J}_0^z$ is acceptable for all $z \in \mathcal{Z}_\varepsilon$. This and the acceptability (by definition) of $\hat{\mathcal{J}}_z$ imply that on $\mathcal{F}^c$, for all $z \in \mathcal{Z}_\varepsilon$,

$$\begin{aligned}\Big|\Delta_{\mathcal{J}_0^z}\hat{g}_{\mathcal{J}_0^z*\hat{\mathcal{J}}_z}(z)\Big| &\leq \mathbf{TH}_\varepsilon\big(\mathcal{J}_0^z, \hat{\mathcal{J}}_z\big), \\ \Big|\Delta_{\hat{\mathcal{J}}_z}\hat{g}_{\hat{\mathcal{J}}_z*\mathcal{J}_0^z}(z)\Big| &\leq \mathbf{TH}_\varepsilon\big(\hat{\mathcal{J}}_z, \mathcal{J}_0^z\big).\end{aligned} \tag{61}$$

This, the definition of the threshold $\mathbf{TH}_\varepsilon$ and the fact that $\|K_{\mathcal{J}}\|_2 = \|\mathsf{K}_{(\mathcal{A},\lambda_\varepsilon(\mathcal{A}))}\|_2$ for all $\mathcal{J} \in \mathfrak{J}_{\mathrm{grid}}$ yield that on $\mathcal{F}^c$, for all $z \in \mathcal{Z}_\varepsilon$,

$$\begin{aligned}\Big|\hat{g}_{\hat{\mathcal{J}}_z}(z) - g(z)\Big| &\leq 4C(p,d)K_{\mathcal{A}}^* \, \|\mathsf{K}_{(\mathcal{A},\lambda_\varepsilon(\mathcal{A}))}\|_2 \, \varepsilon\sqrt{\ln(1/\varepsilon)} + \Big|\hat{g}_{\mathcal{J}_0^z}(z) - g(z)\Big| \\ &= 4C(p,d)c_{11}^{-1}K_{\mathcal{A}}^* \, \lambda_\varepsilon(\mathcal{A}) + \Big|\hat{g}_{\mathcal{J}_0^z}(z) - g(z)\Big|.\end{aligned} \tag{62}$$

We combine (59) and (62) to get, with some constants $c_{14} - c_{16}$ independent of ε,

$$\begin{aligned}\sup_{g\in\mathbb{H}(\mathcal{A},\mathcal{L})} \mathbb{E}_g\Big(|\bar{g}_\varepsilon^* - g|_\infty^p \mathbb{I}\{\mathcal{F}^c\}\Big) &\leq c_{14}(K_{\mathcal{A}}^* \, \lambda_\varepsilon(\mathcal{A}))^p + c_{15}\phi_\varepsilon^p(\gamma,\beta) \\ &\leq c_{16}(K_{\mathcal{A}}^*\phi_\varepsilon(\gamma,\beta))^p.\end{aligned} \tag{63}$$

Theorem 2 follows now from (58) and (63).

A Proofs of auxiliary results

A.1 Proof of Proposition 1

1°. *Preliminary remarks.*
For any $\mathcal{J} \in \mathfrak{J}$ and any $x \in [-1,1]^d$ we may write

$$\begin{aligned}
&\big[\Delta_{\mathcal{J}} K_{\mathcal{J}*\mathcal{J}_0^x} * g\big](x) = \big[K_{\mathcal{J}*\mathcal{J}_0^x} * g\big](x) - \big[K_{\mathcal{J}} * g\big](x) \qquad (64)\\
&= \int \left(\int K_{\mathcal{J}}(y-x) K_{\mathcal{J}_0^x}(t-y) dy \right) g(t) dt - \big[K_{\mathcal{J}} * g\big](x)\\
&= \int K_{\mathcal{J}}(y-x) \left(\int K_{\mathcal{J}_0^x}(t-y) g(t) dt \right) dy - \big[K_{\mathcal{J}} * g\big](x)\\
&= \int K_{\mathcal{J}}(y-x) g(y) dy - \big[K_{\mathcal{J}} * g\big](x)\\
&\quad + \int K_{\mathcal{J}}(y-x) \left(\int K_{\mathcal{J}_0^x}(t-y) \big[g(t) - g(y)\big] dt \right) dy\\
&= \int K_{\mathcal{J}}(y-x) \left(\int K_{\mathcal{J}_0^x}(t-y) \big[g(t) - g(y)\big] dt \right) dy\\
&= \int K_{\mathcal{J}}(v) \left[\int K_{\mathcal{J}_0^x}(z) \big(g(z+v+x) - g(v+x)\big) dz \right] dv\\
&= \int \mathsf{K}_{(\mathcal{A},\lambda)}(M_\vartheta^T v) \int \mathsf{K}_{(\mathcal{A}_0,\lambda_0)}(M_{\vartheta^x}^T z) (g(z+v+x) - g(v+x)) dz dv.
\end{aligned}$$

Define $G_x(\cdot) = G(\cdot + x)$ and $f_x(\cdot) = f(\cdot + G(x))$. Then $g(z+v+x) = f(G_x(z+v))$ and $g(v+x) = f(G_x(v))$. Note that, for all $x \in [-1,1]^d$,

$$G_x \in \mathbb{H}_d(\beta, L_2), \quad f_x \in \mathbb{H}_1(\gamma, L_1). \qquad (65)$$

If $1 < \gamma \le 2$, the second property in (65) implies

$$f_x' \in \mathbb{H}_1(\gamma - 1, 2L_1). \qquad (66)$$

In the case where $1 < \beta \le 2$, for all $u \in \mathbb{R}^d$, $x \in [-1,1]^d$ we define $\tilde{G}_x(u) = G_x(u) - G_x(0) - [\nabla G_x(0)]^T u$. In view of (65), for all $x \in [-1,1]^d$ we have

$$\|\nabla \tilde{G}_x(u)\| \le 2L_2, \quad \forall\, u \in \mathbb{R}^d, \qquad (67)$$

$$\begin{aligned}
&\Big|\tilde{G}_x(t) - \tilde{G}_x(u) - [\nabla \tilde{G}_x(u)]^T (t-u)\Big| \le L_2 \|t-u\|^\beta, \quad \forall\, t, u \in \mathbb{R}^d, \qquad (68)\\
&\Rightarrow \quad |\tilde{G}_x(u)| \le L_2 \|u\|^\beta, \quad u \in \mathbb{R}^d.
\end{aligned}$$

It follows from the definition of $\mathsf{K}_{(\mathcal{A},\lambda)}$ and Lemmas 1 – 4 that

$$\int \|v\|^{\gamma\beta} \big|\mathsf{K}_{(\mathcal{A},\lambda)}(v)\big| dv \le c_6' \lambda^{\frac{\gamma\beta}{ab}}, \quad \forall\, \mathcal{A} \in (0,2]^2,\ \lambda > 0, \qquad (69)$$

where $c_6' > 0$ is a constant depending only on $\mathcal{L}$ and d. Furthermore, for any $\mathcal{A} = (a,b) \in (0,2]^2$ and any $\lambda \le 1$ the support of $\mathsf{K}_{(\mathcal{A},\lambda)}$ is contained in a ball $\{u \in \mathbb{R}^d : \|u\| \le c_K \lambda^{\frac{1}{ab}}\}$ where the constant $c_K > 0$ depends only on d. Therefore,

$$\mathsf{K}_{(\mathcal{A},\lambda)}\left(M_\vartheta^T u\right) = 0, \quad \forall\, u, \vartheta \in \mathbb{R}^d : \|u\| > c_K \lambda^{\frac{1}{ab}},\ \|\vartheta\| = 1. \qquad (70)$$

2°. *Proof for the zone of combined local model:* $1 < \gamma \le \beta \le 2$.

Using (65) and the Taylor expansion for G_x we obtain, for all $x \in [-1,1]^d$, $z, v \in \mathbb{R}^d$,

$$\begin{aligned} g(z+v+x) &= f\Big(G_x(0) + \big[\nabla G_x(0)\big]^T(z+v) + \tilde{G}_x(z+v)\Big) \\ &= f_x\Big(\big[\nabla G_x(0)\big]^T(z+v) + \tilde{G}_x(z+v)\Big). \end{aligned} \tag{71}$$

Note that, by definition, $\nabla G_x(0) = \nabla G(x) = \vartheta_0^x \|\nabla G(x)\|$. Set $\nabla G_* = \vartheta^x \|\nabla G(x)\|$ and define

$$g_*(z+v+x) = f_x\Big(\big[\nabla G_*\big]^T(z+v) + \tilde{G}_x(z+v)\Big).$$

We now approximate $g(z+v+x)$ by $g_*(z+v+x)$ in the last line of (64). In view of (70), it suffices to consider there only the values z, v satisfying $\|z\|, \|v\| \le c_K$. For such z, v and all $x \in [-1,1]^d$, the condition $\|\vartheta_0^x - \vartheta^x\| \le \varepsilon^2$ and (65) imply

$$|g(z+v+x) - g_*(z+v+x)| \;\le\; 2c_K L_1 \|\nabla G(x)\| \varepsilon^2 \le 2c_K L_1 L_2\, \varepsilon^2. \tag{72}$$

Using (65) – (68), the Taylor expansion for f_x and (66), we get that for all $x \in [-1,1]^d$, $z, v \in \mathbb{R}^d$ the following representation holds:

$$\begin{aligned} g_*(z+v+x) &= f_x\Big(\big[\nabla G_*\big]^T(z+v)\Big) \\ &\quad + f_x'\Big(\big[\nabla G_*\big]^T(z+v)\Big)\tilde{G}_x(z+v) + B_{x,1}(z,v)\|z+v\|^{\gamma\beta} \\ &= f_x\Big(\big[\nabla G_*\big]^T(z+v)\Big) \\ &\quad + \Big[f_x'\Big(\big[\nabla G_*\big]^T(z+v)\Big) - f_x'\Big(\big[\nabla G_*\big]^T v\Big)\Big] \\ &\qquad \times \Big(\tilde{G}_x(v) + \big[\nabla \tilde{G}_x(v)\big]^T z\Big) \\ &\quad + f_x'\big(\big[\nabla G_*\big]^T v\big)\Big(\tilde{G}_x(z+v) - \tilde{G}_x(v)\Big) \\ &\quad + f_x'\big(\big[\nabla G_*\big]^T v\big)\tilde{G}_x(v) \\ &\quad + B_{x,2}(z,v)\Big|\big[\nabla G_*\big]^T z\Big|^{\gamma-1}\|z\|^\beta + B_{x,1}(z,v)\|z+v\|^{\gamma\beta}, \end{aligned} \tag{73}$$

where, for all $x \in [-1,1]^d$, $z, v \in \mathbb{R}^d$, $B_{x,1}(\cdot,\cdot)$ and $B_{x,2}(\cdot,\cdot)$ are functions satisfying

$$\big|B_{x,1}(z,v)\big| \le L_1 L_2^\gamma, \quad \big|B_{x,2}(z,v)\big| \le 2L_1 L_2. \tag{74}$$

Putting $z = 0$ in (73) we obtain

$$g_*(v+x) = f_x\Big(\big[\nabla G_*\big]^T v\Big) + f_x'\Big(\big[\nabla G_*\big]^T v\Big)\tilde{G}_x(v) + B_{x,1}(0,v)\|v\|^{\gamma\beta}. \tag{75}$$

From (73) and (75) we get, for all $x \in [-1,1]^d$, $z, v \in \mathbb{R}^d$,

$$\begin{aligned} &g_*(z+v+x) - g_*(v+x) \\ &= f_x\Big(\big[\nabla G_*\big]^T(z+v)\Big) - f_x\Big(\big[\nabla G_*\big]^T v\Big) \\ &\quad + \Big[f_x'\Big(\big[\nabla G_*\big]^T(z+v)\Big) - f_x'\Big(\big[\nabla G_*\big]^T v\Big)\Big]\Big(\tilde{G}_x(v) + \big[\nabla \tilde{G}_x(v)\big]^T z\Big) \\ &\quad + f_x'\Big(\big[\nabla G_*\big]^T v\Big)\Big(\tilde{G}_x(z+v) - \tilde{G}_x(v)\Big) \\ &\quad + B_{x,2}(z,v)\Big|\big[\nabla G_*\big]^T z\Big|^{\gamma-1}\|z\|^\beta + B_{x,1}(z,v)\|z+v\|^{\gamma\beta} \\ &\quad - B_{x,1}(0,v)\|v\|^{\gamma\beta}. \end{aligned} \tag{76}$$

Put $u = M_{\vartheta^x}^T v,\ s = M_{\vartheta^x}^T z$. We get from (76) that

$$\begin{aligned} & g_*(M_{\vartheta^x} s + M_{\vartheta^x} u + x) - g_*(M_{\vartheta^x} u + x) \\ & = \big(\tilde f_x(s_1+u_1) - \tilde f_x(u_1)\big) \\ & \quad + A_{u,x}(s_1)\Big(\overline{G}_x(u) + \big[\nabla \overline{G}_x(u)\big]^T s\Big) \\ & \quad + f_x'\big(\|\nabla G(x)\| u_1\big)\Big(\overline{G}_x(s+u) - \overline{G}_x(u)\Big) + \tilde B_{x,2}(s,u)|s_1|^{\gamma-1}\|s\|^\beta \\ & \quad + \tilde B_{x,1}(s,u)\|s+u\|^{\gamma\beta} - \tilde B_{x,1}(0,u)\|u\|^{\gamma\beta}, \end{aligned} \tag{77}$$

where s_1 and u_1 are the first components of $s \in \mathbb{R}^d$ and $u \in \mathbb{R}^d$ respectively,

$$\begin{aligned} \tilde f_x(u_1) &= f_x(\|\nabla G(x)\| u_1),\ \overline{G}_x(u) = \tilde G_x(M_{\vartheta^x} u), \\ \tilde B_{x,1}(s,u) &= B_{x,1}(M_{\vartheta^x} s, M_{\vartheta^x} u) \\ \tilde B_{x,2}(s,u) &= \|\nabla G(x)\|^{\gamma-1} B_{x,2}(M_{\vartheta^x} s, M_{\vartheta^x} u), \end{aligned}$$

and

$$A_{u,x}(s_1) = f_x'\big(\|\nabla G(x)\|(s_1+u_1)\big) - f_x'\big(\|\nabla G(x)\| u_1\big).$$

It is easy to see that inequalities (67) and (68) remain valid with $\overline{G}_x$ in place of $\tilde G_x$.

Now for all $x \in [-1,1]^d, s, u \in \mathbb{R}^d$ we introduce

$$\begin{aligned} q_{u,x}(s_1) &= \big(\tilde f_x(s_1+u_1) - \tilde f_x(u_1)\big) + A_{u,x}(s_1)\big(\overline{G}_x(u) + \big[\nabla \overline{G}_x(u)\big]^T \vartheta^x s_1\big) \\ &\quad + f_x'\big(\|\nabla G(x)\| u_1\big)\big[\nabla \overline{G}_x(u)\big]^T \vartheta^x s_1, \\ p_{u,x}(s) &= f_x'\big(\|\nabla G(x)\| u_1\big)\Big(\overline{G}_x(s+u) - \overline{G}_x(u) - \big[\nabla \overline{G}_x(u)\big]^T s\Big), \\ B^{u,x}(s) &= \tilde B_{x,2}(s,u), \quad Q_{u,x}(s) = q_{u,x}(s_1) + p_{u,x}(s) + \tilde B_{x,2}(s,u)|s_1|^{\gamma-1}\|s\|^\beta, \\ P_{u,x}(s) &= f_x'\big(\|\nabla G(x)\|(s_1+u_1)\big)\big[\nabla \overline{G}_x(u)\big]^T s_\perp \end{aligned}$$

where $s_\perp = s - s_1 \vartheta^x$. With this notation (77) can be written as

$$\begin{aligned} g_*(M_{\vartheta^x} s + M_{\vartheta^x} u + x) - g_*(M_{\vartheta^x} u + x) = Q_{u,x}(s) + P_{u,x}(s) \\ + \tilde B_{x,1}(s,u)\|s+u\|^{\gamma\beta} - \tilde B_{x,1}(0,u)\|u\|^{\gamma\beta}. \end{aligned} \tag{78}$$

We now prove that, for all $x \in [-1,1]^d$ and all $u \in \mathbb{R}^d$ such that $\|u\| \le c_K \lambda^{\frac{1}{ab}}$ (cf. (70)), where $\lambda^{\frac{1}{ab}} \le 2\lambda_0^{\frac{1}{\gamma\beta}}$, the triplet $(q_{u,x}, p_{u,x}, B^{u,x})$ belongs to the set $\mathfrak{B}(\mathcal{A}_0, \lambda_0)$ (cf. definition before Lemma 3), and thus Lemmas 3 or 4 can be applied. We need to check (31) – (33).

Checking (31). In view of (65) we have

$$|\tilde f_x(s_1+u_1) - \tilde f_x(u_1) - \tilde f_x'(u_1)s_1| \le L_1 L_2 |s_1|^\gamma.$$

Therefore,

$$\left|\frac{1}{2\lambda_0^{1/\gamma}} \int_{-\lambda_0^{1/\gamma}}^{\lambda_0^{1/\gamma}} \Big(\tilde f_x(s_1+u_1) - \tilde f_x(u_1)\Big)\, ds_1\right| \le \frac{L_1 L_2}{2\lambda_0^{1/\gamma}} \int_{-\lambda_0^{1/\gamma}}^{\lambda_0^{1/\gamma}} |s_1|^\gamma\, ds_1 \le \frac{L_1 L_2}{2}\lambda_0. \tag{79}$$

Next, remark that (66) implies $|A_{u,x}(s_1)| \le 2L_1 L_2^{\gamma-1}|s_1|^{\gamma-1}$. Furthermore, (68) with $\overline{G}_x$ in place of $\tilde G_x$ yields $|\overline{G}_x(u)| \le L_2\|u\|^\beta$. Now, $q_{u,x}(0) = 0$ and using these remarks, (79) and

(67) we get, for $\|u\| \le c_K \lambda^{\frac{1}{ab}}$, $\lambda^{\frac{1}{ab}} \le 2\lambda_0^{\frac{1}{\gamma\beta}}$,

$$\begin{aligned}
&\left| \frac{1}{2\lambda_0^{1/\gamma}} \int_{-\lambda_0^{1/\gamma}}^{\lambda_0^{1/\gamma}} q_{u,x}(s_1) ds_1 \right| \qquad (80)\\
&\quad \le \frac{L_1 L_2}{2} \lambda_0 + \frac{1}{2\lambda_0^{1/\gamma}} \int_{-\lambda_0^{1/\gamma}}^{\lambda_0^{1/\gamma}} |A_{u,x}(s_1)| \left(|\overline{G}_x(u)| + \|\nabla \overline{G}_x(u)\| |s_1| \right) ds_1 \\
&\quad \le \frac{L_1 L_2}{2} \lambda_0 + 2 L_1 L_2^\gamma \left(\frac{1}{\gamma} \lambda_0^{(\gamma-1)/\gamma} \|u\|^\beta + \frac{2}{\gamma+1} \lambda_0 \right) \\
&\quad \le \left[\frac{L_1 L_2}{2} + 2 L_1 L_2^\gamma \left(\frac{(2c_K)^\beta}{\gamma} + \frac{2}{\gamma+1} \right) \right] \lambda_0 \le c_3 \lambda_0
\end{aligned}$$

where the constant c_3 depends only on $\mathcal{L}$ and d. It can be taken as a maximum of the last expression in square brackets over $(\gamma, \beta) \in [1,2]^2$.

Checking (32) and (33). It suffices to note that, for all $x \in [-1,1]^d$, the first property in (68) with $\overline{G}_x$ in place of $\tilde{G}_x$ and the second property in (65) yield

$$\begin{aligned}
\left| p_{u,x}(s') - p_{u,x}(s) - [\nabla p_{u,x}(s)]^T (s'-s) \right| &\le \left| f_x'\big(\|\nabla G(x)\| u_1\big) \right| L_2 \|s'-s\|^\beta \\
&\le L_1 L_2 \|s'-s\|^\beta, \quad \forall\, s, s' \in \mathbb{R}^d.
\end{aligned}$$

This proves (32) with $b = \beta$ and $L = L_1 L_2$. Finally, (33) with $B = B^{u,x}$, $c_4 = 2L_1 L_2^\gamma$ follows from (74).

We are now in a position to apply Lemmas 3 and 4. We demonstrate this, for example, for Lemma 4. Take there $q = q_{u,x}$, $p = p_{u,x}$, $B = B^{u,x}$ for any $\|u\| \le c_K \lambda^{\frac{1}{ab}}$ and $x \in [-1,1]^d$. Since $Q_{u,x}(0) = 0$, the result (39) of Lemma 4 yields

$$\left| \int \mathsf{K}_{(\mathcal{A}_0,\lambda_0)}(s) Q_{u,x}(s) ds \right| \le c_5 \lambda_0 \qquad (81)$$

where c_5 depends only on $\mathcal{L}$ and d. Furthermore, by construction the kernel $\mathsf{K}_{(\mathcal{A}_0,\lambda_0)}$ is symmetric, i.e., $\mathsf{K}_{(\mathcal{A}_0,\lambda_0)}(s) = \mathsf{K}_{(\mathcal{A}_0,\lambda_0)}(-s)$ and hence

$$\int \mathsf{K}_{(\mathcal{A}_0,\lambda_0)}(s) P_{u,x}(s) ds = 0. \qquad (82)$$

Next, using (74) we find

$$\left| \tilde{B}_{x,1}(s,u) \|s+u\|^{\gamma\beta} - \tilde{B}_{x,1}(0,u) \|u\|^{\gamma\beta} \right| \le 2^{\gamma\beta} L_1 L_2^\gamma \Big(\|s\|^{\gamma\beta} + \|u\|^{\gamma\beta} \Big).$$

Combining this inequality and (81) – (82) with (78) we get, for all $x \in [-1,1]^d, u \in \mathbb{R}^d$,

$$\begin{aligned}
&\left| \int \mathsf{K}_{(\mathcal{A}_0,\lambda_0)}(s) (g_*(M_{\vartheta^x} s + M_{\vartheta^x} u + x) - g_*(M_{\vartheta^x} u + x)) ds \right| \\
\le\ & c_5 \lambda_0 + 2^{\gamma\beta} L_1 L_2^\gamma \left[\int \left| \mathsf{K}_{(\mathcal{A}_0,\lambda_0)}(s) \right| \|s\|^{\gamma\beta} ds + \left\| \mathsf{K}_{(\mathcal{A}_0,\lambda_0)} \right\|_1 \|u\|^{\gamma\beta} \right].
\end{aligned}$$

We finally get (43) from this inequality invoking (69), (64), (72) and the condition $\lambda^{\frac{1}{ab}} \le 2\lambda_0^{\frac{1}{\gamma\beta}}$ and recalling that $\left\| \mathsf{K}_{(\mathcal{A},\lambda)} \right\|_1 = \left\| K_{\mathcal{J}} \right\|_1$ for all $\mathcal{A} \in (0,2]^2$, $\lambda > 0$, and $\left\| \mathsf{K}_{(\mathcal{A}_0,\lambda_0)} \right\|_1 = \left\| K_{\mathcal{J}_0^x} \right\|_1$.

3°. *Proof of (43) for the local single index zone:* $\gamma \le 1,\ 1 < \beta \le 2$.
Using (68) and the second property in (65), for all $z, v \in \mathbb{R}^d$, $x \in [-1,1]^d$ we may write

$$g_*(z+v+x) = f_x\Big(\big[\nabla G_*\big]^T(z+v)\Big) + B_{x,1}(z,v)\|z+v\|^{\gamma\beta},$$

where $B_{x,1}$ satisfies (74). This can be viewed as a simplified version of (73). Following almost the same argument as in 2° (the main difference is that now we drop all the terms containing f'_x and $B_{x,2}$) and applying Lemma 2 we obtain (43).

4°. *Proof of (43) for the zone of slow rate:* $(\gamma,\beta) \in (0,1]^2$.
Using the Hölder condition on f and G_x we obtain, for all $z, v \in \mathbb{R}^d$, $x \in [-1,1]^d$,

$$g(z+v+x) \equiv f\big(G_x(z+v)\big) = f\big(G_x(0)\big) + B_{x,1}(z,v)\|z+v\|^{\gamma\beta}$$

where $B_{x,1}$ satisfies (74). Now, (43) easily follows from this relation, (64), (69), the definition of $\mathsf{K}_{(\mathcal{A}_0,\lambda_0)}$ for the zone of slow rate and the condition $\lambda^{\frac{1}{ab}} \le 2\lambda_0^{\frac{1}{\gamma\beta}}$.

5°. *Proof of (43) for the zone of inactive structure:* $1 < \beta \le \gamma \le 2$.
Since $f \in \mathbb{H}_1(\gamma, L_1)$ and $\|\nabla G_x(\cdot)\| \le L_2$, for all $z, v \in \mathbb{R}^d$, $x \in [-1,1]^d$ we may write

$$\begin{aligned}
f\big(G_x(z+v)\big) &= f\big(G_x(v)\big) + f'\big(G_x(v)\big)\big(G_x(z+v) - G_x(v)\big) + B_{x,1}(z,v)\|z\|^\gamma \\
&= f\big(G_x(v)\big) + f'\big(G_x(v)\big)\big(G_x(z+v) - G_x(v) - \big[\nabla G_x(v)\big]^T z\big) \\
&\quad + f'\big(G_x(v)\big)\big[\nabla G_x(v)\big]^T z + B_{x,1}(z,v)\|z\|^\gamma \\
&= f\big(G_x(v)\big) + f'\big(G_x(v)\big)\big[\nabla G_x(v)\big]^T z + B_{x,2}(z,v)\|z\|^\beta + B_{x,1}(z,v)\|z\|^\gamma
\end{aligned}$$

where $B_{x,1}$ satisfies (74) and $|B_{x,2}(\cdot,\cdot)| \le L_1 L_2$. Since the kernel $\mathsf{K}_{(\mathcal{A}_0,\lambda_0)}$ is symmetric,

$$\int \mathsf{K}_{(\mathcal{A}_0,\lambda_0)}(M_{\vartheta^x}^T z) f'\big(G_x(v)\big)\big[\nabla G_x(v)\big]^T z\, dz = 0.$$

Now, (43) easily follows from these relations, (64), the definition of $\mathsf{K}_{(\mathcal{A}_0,\lambda_0)}$ for the zone of inactive structure and the condition $\lambda \le 1$.

6°. *Proof of (44).* For a function $K \in L_2(\mathbb{R}^d)$, let us denote by $\widehat{K}$ its Fourier transform. Using Parceval's identity we obtain, for any $\mathcal{J}, \mathcal{J}' \in \mathfrak{J}$,

$$\begin{aligned}
\|\Delta_{\mathcal{J}'} K_{\mathcal{J}'*\mathcal{J}}\|_2 &= \frac{1}{\sqrt{2\pi}}\|\widehat{\Delta_{\mathcal{J}'} K_{\mathcal{J}'*\mathcal{J}}}\|_2 = \frac{1}{\sqrt{2\pi}}\|\big(\widehat{K}_{\mathcal{J}} - 1\big)\widehat{K}_{\mathcal{J}'}\|_2 \\
&\le \frac{1}{\sqrt{2\pi}}\Big(\|\widehat{K}_{\mathcal{J}}\|_\infty + 1\Big)\ \|\widehat{K}_{\mathcal{J}'}\|_2 \le \big(\|K_{\mathcal{J}}\|_1 + 1\big)\|K_{\mathcal{J}'}\|_2.
\end{aligned}$$

Since $\int K_{\mathcal{J}'} = 1$, this proves (44).

A.2 Proof of Lemma 3

First, note that some cases are trivial because the number r of steps of the kernel construction is bounded by 3. In fact, if $(\rho+1)\rho < (b-a)/a$ and $V(\lambda) \le \ln\Big(\frac{\sqrt{5}+1}{2}\Big)$ we have $r \le 3$ by definition. If $(\rho+1)\rho \ge (b-a)/a$ we use the kernel as in Lemma 3. But for this kernel the condition $(\rho+1)\rho \ge (b-a)/a$ implies that, again, $r \le 3$.

So, we will treat only the remaining case where $(\rho+1)\rho < (b-a)/a$ and $V(\lambda) > \ln\Big(\frac{\sqrt{5}+1}{2}\Big)$. The last inequality implies that $r > 3$.

Note that, by definition, $\alpha < \frac{1}{2}\ln\left(\frac{\sqrt{5}+1}{2}\right)$. Further, for $r \geq 3$ we have also the lower bound: $\alpha \geq \frac{1}{4}\ln\left(\frac{\sqrt{5}+1}{2}\right)$. Thus for $r \geq 3$,

$$0.786 \leq \left(\frac{\sqrt{5}+1}{2}\right)^{-1/2} < e^{-\alpha} \leq \left(\frac{\sqrt{5}+1}{2}\right)^{-1/4} \leq 0.887. \tag{83}$$

1°. *Proof of (39).* From the definition of $\mathsf{K}_{(\mathcal{A},\lambda)}$ we find

$$\begin{aligned}[\mathsf{K}_{(\mathcal{A},\lambda)} * q](0) &= 2^{-d}\sum_{i=1}^{r}\int \Lambda_i(|y|)q(y_1)dy = 2^{-d}\int \Lambda_1(|y|)q(y_1)dy \\ &= \frac{1}{u_1}\int \frac{q(y_1)+q(-y_1)}{2}\mathbb{I}_{[0,u_1]}(y_1)dy_1\end{aligned}$$

where $u_1 = \lambda^{1/a}$. This and (31) imply

$$\left|\big[\mathsf{K}_{(\mathcal{A},\lambda)} * q\big](0) - q(0)\right| = \left|\big(2\lambda^{1/a}\big)^{-1}\int_{-\lambda^{1/a}}^{\lambda^{1/a}} q(y_1)dy_1 - q(0)\right| \leq c_3\lambda. \tag{84}$$

We now obtain a similar bound for $\left|\big[\mathsf{K}_{(\mathcal{A},\lambda)} * p\big](0) - p(0)\right|$. Note that, in view of (32), for all $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$ we have

$$p(z) = \tilde{p}(z) + z_1\frac{\partial p}{\partial z_1}(0, z_2, \ldots, z_d) + B_1(z)z_1^b, \tag{85}$$

where $\tilde{p}(z) = p(0, z_2, \ldots, z_d)$ and $\sup_{z\in\mathbb{R}^d}|B_1(z)| \leq L$. For the same reason, for all $z_{(d-1)} \triangleq (0, z_2, \ldots, z_d)$ we have

$$\tilde{p}(z) = \tilde{p}(0) + [\nabla\tilde{p}(0)]^T z_{(d-1)} + B_2(z_{(d-1)})\|z_{(d-1)}\|^b, \tag{86}$$

where as previously $|B_2(\cdot)| \leq L$. Combining (85) and (86) and taking into account that the function $\mathsf{K}_{(\mathcal{A},\lambda)}$ is symmetric, $\int \mathsf{K}_{(\mathcal{A},\lambda)} = 1$ and $\tilde{p}(0) = p(0)$ we get

$$\left|\big[\mathsf{K}_{(\mathcal{A},\lambda)} * p\big](0) - p(0)\right| = \left|\int \mathsf{K}_{(\mathcal{A},\lambda)}(z)\big(B_1(z)z_1^b + B_2(z_{(d-1)})\|z_{(d-1)}\|^b\big)dz\right| \tag{87}$$

Now

$$\begin{aligned}&\left|\int \mathsf{K}_{(\mathcal{A},\lambda)}(z)B_2(z_{(d-1)})\|z_{(d-1)}\|^b dz\right| \qquad (88)\\ &= \left|\big(2(v_1-v_2)\big)^{1-d}\int B_2(z_{(d-1)})\|z_{(d-1)}\|^b\mathbb{I}_{[v_2,v_1]^{d-1}}(|z_{(d-1)}|)dz_{(d-1)}\right.\\ &\quad + \sum_{i=1}^{r-1}\left[\big(2(v_i-v_{i+1})\big)^{1-d}\int B_2(z_{(d-1)})\|z_{(d-1)}\|^b\mathbb{I}_{[v_{i+1},v_i]^{d-1}}(|z_{(d-1)}|)dz_{(d-1)}\right.\\ &\quad\quad \left.\left. -\big(2(v_{i-1}-v_i)\big)^{1-d}\int B_2(z_{(d-1)})\|z_{(d-1)}\|^b\mathbb{I}_{[v_i,v_{i-1}]^{d-1}}(|z_{(d-1)}|)dz_{(d-1)}\right]\right|\\ &\leq \big(2v_r\big)^{1-d}\int \big|B_2(z_{(d-1)})\big|\,\|z_{(d-1)}\|^b\mathbb{I}_{[0,v_r]^{d-1}}(|z_{(d-1)}|)dz_{(d-1)}\\ &= \big(\lambda^{1/b}\big)^{1-d}\int \big|B_2(z_{(d-1)})\big|\,\|z_{(d-1)}\|^b\mathbb{I}_{[0,\lambda^{1/b}]^{d-1}}(|z_{(d-1)}|)dz_{(d-1)}\\ &\leq 2^{d-1}d^{\frac{b}{2}}L\lambda \leq 2^{d-1}dL\lambda\end{aligned}$$

where $|z_{(d-1)}| = (|z_2|, \dots, |z_d|)$. Further, note that $v \ge u \ge 1$ implies $e^{\frac{v}{u}} \le e^v/u$ (in fact, $v(1-1/u) \ge u - 1 \ge \ln u$). Using this remark and the fact that $\frac{b}{a-1} > 1$ we find

$$\begin{aligned} u_i &= \lambda^{\frac{1}{a}} \exp\left(\frac{b}{a-1}\exp\left(\alpha(i-1)\right)\right) = \lambda^{\frac{1}{a}} \exp\left(\frac{b}{a-1}\exp\left(\alpha i\right) e^{-\alpha}\right) \\ &\le u_{i+1} e^{-\alpha}, \quad i = 1, \dots, r-1, \end{aligned} \tag{89}$$

and therefore $u_i/u_r \le e^{\alpha(i-r)}$. This and the equality $u_r = \lambda^{\frac{1}{b}}$ allow us to get

$$\begin{aligned} \left|\int \mathsf{K}_{(\mathcal{A},\lambda)}(z) B_1(z) z_1^b dz\right| &\le L \int \left|\mathsf{K}_{(\mathcal{A},\lambda)}(z)\right| |z_1|^b dz \qquad (90) \\ &= \frac{L}{u_1}\int z_1^b \mathbb{I}_{[0,u_1]}(z_1) dz_1 + \sum_{i=2}^{r} \frac{2L}{u_i - u_{i-1}} \int z_1^b \mathbb{I}_{[u_{i-1},u_i]}(z_1) dz_1 \\ &\le 2L \sum_{i=1}^{r} u_i^b \le 2L\lambda \sum_{i=1}^{r} \left(\frac{u_i}{u_r}\right)^b \le 2\lambda L \sum_{l=0}^{\infty} e^{-\alpha l} = 2\lambda L (1 - e^{-\alpha})^{-1}. \end{aligned}$$

From (87), (88) and (90) we get

$$\left|\left[\mathsf{K}_{(\mathcal{A},\lambda)} * p\right](0) - p(0)\right| \le \lambda L \left[2^{d-1} d + 2(1 - e^{-\alpha})^{-1}\right]. \tag{91}$$

We now estimate the value $\left|\int \mathsf{K}_{(\mathcal{A},\lambda)}(y) B(y) y_1^{a-1} \|y\|^b dy\right|$. In view of (38),

$$\begin{aligned} u_1^{a-1} v_1^b &\le \lambda \exp\{b - \nu b e^{\alpha}\} \le \lambda \exp\{(1-\nu) b\}, \\ u_i^{a-1} v_i^b &\le u_i^{a-1} v_{i-1}^b = \lambda \exp\Big\{(1-\nu) b \exp(\alpha(i-1))\Big\}, \quad i = 2, \dots, r. \end{aligned} \tag{92}$$

Using (92), we get similarly to (90):

$$\begin{aligned} \Big|\int \mathsf{K}_{(\mathcal{A},\lambda)}(y) B(y) y_1^{a-1} \|y\|^b dy\Big| &\le c_4 \int \Big|\mathsf{K}_{(\mathcal{A},\lambda)}(y)\Big| |y_1|^{a-1} \sum_{j=1}^{d} |y_j|^b dy \qquad (93) \\ &= c_4 \left[\int \Big|\mathsf{K}_{(\mathcal{A},\lambda)}(y)\Big| |y_1|^{a+b-1} dy + \sum_{j=2}^{d} \int \Big|\mathsf{K}_{(\mathcal{A},\lambda)}(y)\Big| |y_1|^{a-1} |y_j|^b dy\right] \\ &\le 2c_4 \left[\sum_{i=1}^{r} u_i^{b+a-1} + d \sum_{i=1}^{r} u_i^{a-1} v_i^b\right] \\ &\le 2c_4 \left[\lambda^{\frac{b+a-1}{b}} \sum_{l=0}^{\infty} e^{-\alpha l (b+a-1)} + \lambda d \sum_{l=0}^{\infty} \exp\Big\{(1-\nu) b \exp(\alpha l)\Big\}\right] \\ &\le 2c_4 \lambda \left[(1 - e^{-\alpha})^{-1} + d\left(1 - e^{(1-\nu)\alpha}\right)^{-1}\right] \end{aligned}$$

where the last inequality holds for $0 < \lambda \le 1$ and we used that $b \exp(\alpha l) \ge \alpha l$, $\nu > 1$. Summing up the results of (84), (91), (93) and taking into account (83) we obtain (39).

2°. *Proof of (40).* In the same way as above we get, for $0 < \lambda \le 1$,

$$\begin{aligned} \int \Big|\mathsf{K}_{(\mathcal{A},\lambda)}(y)\Big| \|y\|^m du &\le d^{\frac{m}{2}} \int \Big|\mathsf{K}_{(\mathcal{A},\lambda)}(y)\Big| \sum_{j=1}^{d} |y_j|^m dy \\ &\le 2d^{\frac{m}{2}} \left[\sum_{i=1}^{r} u_i^m + d \sum_{i=1}^{r} v_i^m\right] \\ &\le C(d) \lambda^{\frac{m}{ab}} \left[(1 - e^{-m\alpha})^{-1} + (1 - e^{m\nu\alpha})^{-1}\right]. \end{aligned}$$

Here and in what follows use the same notation $C(d)$ for possibly different positive constants depending only on d.

3°. *Proof of (41).* Since $\nu < 2 < \frac{b}{b-a}$ we have, for $0 < \lambda \le 1$,

$$v_{r-1} \triangleq \lambda^{\frac{1}{ab}} \exp\Big\{ -\nu \exp(\alpha(r-1))\Big\} = \lambda^{\frac{1}{ab}+\nu\frac{(a-1)(b-a)}{ab^2}} \ge \lambda^{\frac{1}{b}}.$$

By the definition of v_r this implies that $v_{r-1} - v_r \ge \lambda^{\frac{1}{b}}/2$. Further, as $u_r = \lambda^{\frac{1}{b}}$, in view of (89), we have

$$u_r - u_{r-1} \ge (1 - e^{-\alpha})\lambda^{\frac{1}{b}}.$$

We deduce that

$$\mu_{r,r-1} \ge \mu_{r,r} \ge 2^{1-d}\lambda^{d/b}(1 - e^{-\alpha}). \tag{94}$$

Note that by (89),

$$u_{i+1} - u_i \ge (1 - e^{-\alpha})u_{i+1} \quad \text{for } \ i = 1, \dots, r-1.$$

Also, as $\nu > 1$, it is straightforward to check that

$$v_i - v_{i+1} \ge (1 - e^{-\alpha})v_i \quad \text{for } \ i = 1, \dots, r-2.$$

Thus, we get

$$\mu_{1,1} = u_1(v_1 - v_2)^{d-1} \ge (1 - e^{-\alpha})^{d-1} \exp(-(d-1)\nu e^{\alpha})\, \lambda^{\frac{1}{a}+\frac{d-1}{b}}. \tag{95}$$

Recall that we are considering the case where $\rho(1+\rho) < (b-a)/a$, $1 < a \le b \le 2$, so that $\rho(1+\rho) < 1$, and thus $\rho < \frac{\sqrt{5}-1}{2}$. This and the choice of parameters α, ν combined with (83) implies

$$e^{-\alpha} - \rho\nu \ge \left(\frac{\sqrt{5}+1}{2}\right)^{-1/2} - \rho\nu \ge \left(\frac{\sqrt{5}+1}{2}\right)^{-1/2} - \frac{\sqrt{5}-1}{2}\nu \triangleq \delta \ge 0.0891.$$

Now,

$$\frac{b}{a-1}e^{-\alpha} - (d-1)\nu \ge \frac{\delta b}{a-1} \ge 2\delta.$$

Hence, for $i = 2, \dots, r-1$ we have

$$\begin{aligned} \mu_{i,i-1} &\ge \mu_{i,i} \ge C(d)\lambda^{\frac{1}{a}+\frac{d-1}{ab}} \exp\left\{ \frac{b}{a-1}\exp(\alpha(i-1)) - (d-1)\nu\exp(\alpha i)\right\} \\ &\ge C(d)\lambda^{\frac{1}{a}+\frac{d-1}{ab}} \exp\big\{2\delta\exp(\alpha i)\big\}. \end{aligned} \tag{96}$$

Note that

$$\|\mathsf{K}_{(\mathcal{A},\lambda)}\|_2^2 = \mu_{1,1}^{-1} + \sum_{i=2}^{r}\left(\mu_{i,i-1}^{-1} + \mu_{i,i}^{-1}\right) \le \mu_{1,1}^{-1} + 2\sum_{i=2}^{r}\mu_{i,i}^{-1}. \tag{97}$$

We deduce from (94) – (97) that

$$\|\mathsf{K}_{(\mathcal{A},\lambda)}\|_2^2 \ \le \ C(d)\left(\lambda^{\frac{1}{a}+\frac{d-1}{ab}} + \lambda^{-\frac{d}{b}}\right).$$

This proves the second inequality in (41). The first inequality becomes obvious if we note that $V(\lambda) \le \ln\ln(1/\lambda)$ and so $\|\mathsf{K}_{(\mathcal{A},\lambda)}\|_1 = 2r - 1 \le c_7 \ln\ln(1/\lambda)$, for λ small enough, where c_7 is an absolute constant.

A.3 Proof of Lemma 3

Following the same lines as in the proof of (39) in Lemma 4 we obtain the bound (34) of Lemma 3 with

$$c_5 = C(d)(c_3 + Lr + c_4 r).$$

1°. *Proof of (35).* By definition, $u_r = \lambda^{\frac{1}{b}}$ and for $0 < \lambda \le 1$ we have $u_2 \ge \lambda^{\frac{1}{a}}$, so that $v_1 = \lambda^{\frac{1}{b}} u_2^{-\frac{a-1}{b}} \le \lambda^{\frac{1}{ab}}$. Using these remarks and acting as in the proof of (40) in Lemma 4 we obtain, for $0 < \lambda \le 1$,

$$\begin{aligned}\int \left|\mathsf{K}_{(\mathcal{A},\lambda)}(y)\right| \|y\|^m du &\le 2d^{\frac{m}{2}} \left[\sum_{i=1}^r u_i^m + d\sum_{i=1}^r v_i^m\right] \\ &\le 2d^{\frac{m}{2}} r(u_r^m + dv_1^m) \le C(d) r \lambda^{\frac{m}{ab}}.\end{aligned}$$

2°. *Proof of (36).* Observe that $\alpha_{j+1} - \alpha_j > 0$ for $j = 1, ..., r-1$, so that for $\lambda \to 0$ we have $u_j/u_{j-1} \to \infty$ and $v_{j-1}/v_j \to \infty$. In particular,

$$\mu_{j,j-1} = (u_j - u_{j-1})(v_{j-1} - v_j)^{d-1} \ge \mu_{j,j} = (u_j - u_{j-1})(v_j - v_{j+1})^{d-1} \ge \frac{1}{2} u_j v_j^{d-1}$$

for all λ small enough. Next note that, by definition,

$$\alpha_{r-2} \ge (\alpha_{r-1} - b^{-1})\rho^{-1} \ge \frac{b-a}{ab\rho}.$$

Then $u_2 \le \lambda^{\frac{b-a}{ab\rho}}$ and for λ small enough we get by the definition of ρ:

$$\mu_{1,1} \ge \frac{1}{2} u_1 v_1^{d-1} = \frac{1}{2}\lambda^{\frac{d-1}{b}} u_1 \, u_2^{-\rho} = \frac{1}{2}\lambda^{\frac{d-1}{b}} \lambda^{\frac{1}{a} - \frac{b-a}{ab}} = \frac{1}{2}\lambda^{\frac{d}{b}}.$$

Further, as $u_r = \lambda^{\frac{1}{b}}$ and $v_r = \frac{1}{2}\lambda^{\frac{1}{b}}$, $v_{r+1} = 0$,

$$\mu_{r,r} \ge 2^{-d}\lambda^{\frac{d}{b}}$$

for λ small enough. Next, for $1 < j < r$,

$$\mu_{j,j} \ge \frac{1}{2} u_j v_j^{d-1} = \frac{1}{2}\lambda^{(d-1)/b} u_j \, u_{j+1}^{-\rho}.$$

By the definition of the sequence (α_k),

$$(d-1)/b + \alpha_k - \rho/\alpha_{k-1} = d/b, \quad k = 1, \ldots, r-1.$$

Thus

$$\mu_{j,j} \ge \frac{1}{2}\lambda^{\frac{d-1}{b} + \alpha_{r-j} - \rho\alpha_{r-(j+1)}} = \frac{1}{2}\lambda^{d/b}, \quad j = 2, \ldots, r-1.$$

Substitution of the above bounds into (97) yields

$$\|\mathsf{K}_{(\mathcal{A},\lambda)}\|_2^2 \le C(d)\lambda^{-d/b} r.$$

References

[1] Barron, A. (1993) Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. on Information Theory* **39** 930-945.

[2] Bertin, K. (2004) Asymptotically exact minimax estimation in sup-norm for anisotropic Hölder classes. *Bernoulli* **10** 873-888.

[3] Bertin, K. (2004) *Estimation asymptotiquement exacte en norme sup de fonctions multidimensionnelles*. PhD thesis, Université Paris 6. `http://tel-00008028`

[4] Brown, L. and Low, M. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384-2398.

[5] Donoho, D.L. (1994). Asymptotic minimax risk for sup-norm loss: solution via optimal recovery. *Probability Theory and Related Fields* **99** 145-170.

[6] Golubev, G. (1992) Asymptotically minimax estimation of a regression function in an additive model. *Problems of Information Transmission* **28**, 101-112.

[7] Horowitz, J. and Mammen, E. (2007). Rate-optimal estimation for a general class of non-parametric regression models with unknown link function. *Annals of Statistics*, to appear.

[8] Hristache M., Juditsky A. and Spokoiny V. (2001). Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, **29**(3) 593–623.

[9] Ibragimov, I.A., and Hasminskii, R.Z. (1981). *Statistical Estimation. Asymptotic Theory*, Springer-Verlag. (Originally published in Russian in 1979).

[10] Ibragimov, I.A., and Hasminskii, R.Z. (1982). Bounds for the risks of nonparametric regression estimates. *Theory of Probability and its Applications* **27** 84-99.

[11] Kerkyacharian, G., Lepski, O.V., and Picard, D. (2001) Nonlinear estimation in anisotropic multiindex denoising I. *Probability Theory and Related Fields* **121** 137-170.

[12] Kerkyacharian, G., Lepski, O.V., and Picard, D. (2003) Nonlinear estimation in anisotropic multi-index denoising II. *Theory of Probability and its Applications* **52** 150-171.

[13] Kolmogorov, A.N. (1957) On representation of continuous functions of several variables as superposition of continuous functions of one variable and addition. *Dokl. Akad. Nauk SSSR* **114** 369-373.

[14] Lepski, O. V. (1991). Asymptotically minimax adaptive estimation I: upper bounds. Optimal adaptive estimates. *Theory of Probability and its Applications* **36** 682-697.

[15] Lepski, O. V. (1992a). Asymptotically minimax adaptive estimation II: Statistical models without optimal adaptation. Adaptive estimators. *Theory of Probability and its Applications* **37** 433-468.

[16] Lepski, O. V. (1992b). On problems of adaptive estimation in white Gaussian noise. In *Topics in Nonparametric Estimation*. Advances in Soviet Math., vol. 12 (Khasminskii R. Z., ed.), p. 87-106, Amer. Math. Soc., Providence, R.I.

[17] Lepski, O. V. (1998). Lectures at the *Seminar Paris-Berlin* (Garchy, 1998). Unpiblished.

[18] Lepski, O. V. and Pouet, C. (2007) Hypothesis testing under composite functions alternative. IMA, vol.145, to appear.

[19] Nemirovskii, A. (1985). On nonparametric estimation of smooth regression functions. *Sov. J. Comput. Syst. Sci.* **23** 1-11.

[20] Nussbaum, M. (1986) On nonparametric estimation of a regression function being smooth on a domain in $\mathbb{R}^k$. *Theory of Probability and its Applications* **31** 118-125.

[21] Reiss, M. (2006) Asymptotic equivalence for nonparametric regression with multivariate and random design. `http://arxiv.org/abs/math/0607342`

[22] Sprecher,D.A. (1972) An improvement in the superposition theorem of Kolmogorov. *J. Math. Anal. and Appl.* **38** 208-213.

[23] Stone, C.J. (1982) Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040-1053.

[24] Stone, C.J. (1985) Additive regression and other nonparametric models. *Ann. Statist.* **12** 1285-1297.

[25] Tsybakov, A. B. (2004). *Introduction à l'estimation non-paramétrique,* Springer, Berlin - Heidelberg.